Ana Margarida Pereira Fernandes

Licenciatura em Ciências de Engenharia e Gestão Industrial

Otimização dos processos de Gestão de *Leads*: aplicação de metodologias de *Data Mining*

Dissertação para obtenção do Grau de Mestre em

Engenharia e Gestão Industrial

Orientador: Doutor António Grilo, Professor Associado com Agregação,
Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

Júri:

Presidente: Prof. Doutora Isabel Maria do Nascimento Lopes Nunes

Arguente: Prof. Doutor Pedro Emanuel Botelho Espadinha da Cruz

Vogal: Prof. Doutor António Carlos Bárbara Grilo

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Março 2019

Ana Margarida Pereira Fernandes

Licenciatura em Ciências da Engenharia e Gestão Industrial

Otimização dos processos de Gestão de *Leads*: aplicação de metodologias de *Data Mining*

Dissertação para obtenção do Grau de Mestre em Engenharia e Gestão Industrial

Orientador: Doutor António Grilo, Professor Associado com Agregação,
Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

Júri:

Presidente: Prof. Doutora Isabel Maria do Nascimento Lopes Nunes

Arguente: Prof. Doutor Pedro Emanuel Botelho Espadinha da Cruz

Vogal: Prof. Doutor António Carlos Bárbara Grilo

Março 2019

Otimização dos processos de Gestão de *Leads*: aplicação de metodologias de *Data Mining*

Agradecimentos

Um trabalho desta natureza subentende sem dúvida o envolvimento e a participação de várias pessoas, sem as quais seria difícil concretizá-lo com sucesso. Deste modo, aproveito este espaço para endereçar os meus agradecimentos a quem, de alguma forma, contribuiu para a realização deste trabalho.

Gostaria de começar por agradecer ao Professor António Grilo, que me acompanhou e orientou durante todo este processo, enriquecendo-o com sugestões e críticas construtivas.

À minha família, principalmente aos meus pais, por terem apostado na minha educação e por terem apoiado todas as minhas decisões, mesmo quando estas incluíram atrasar a entrega da dissertação. Não consigo expressar tamanha gratidão por tudo o que têm feito e continuam a fazer por mim.

Um agradecimento às irmãs que me acompanham desde pequenina, que cresceram comigo, e com quem desenvolvi a minha personalidade, capacidade de trabalho, espírito crítico, e principalmente, os meus valores e princípios.

Aos amigos da faculdade, com quem passei tão bons momentos, a estudar ou não. Um agradecimento especial, com muito carinho, à Marisa, à Marta e ao Duarte, que foram uns companheiros de estudo, de trabalhos e de festas incríveis, sem os quais seria quase impossível finalizar com sucesso os 5 anos de curso.

No âmbito da vida profissional, não poderia deixar de agradecer à Joana, à Sofia, ao Rui e à Isabel, que tanto me têm feito crescer enquanto profissional, mas também pessoalmente. Mesmo nos momentos em que o ritmo de trabalho foi alucinante, a vossa motivação e o vosso apoio foi essencial para não desistir, e concluir esta etapa.

E por último, um agradecimento muito especial ao Ricardo, por acreditar em mim e me motivar em todos os momentos. Obrigada pela paciência infindável, pelas palavras de encorajamento nos momentos de maior pressão e por todas as tardes que passaste ao meu lado a ver-me trabalhar (e que não foram poucas).

Resumo

A crescente competitividade do mercado tem pressionado as empresas a adotar estratégias de relação com o cliente mais eficientes e eficazes. Mais especificamente na aquisição de novos clientes, as abordagens de *marketing* direto têm vindo a ganhar um papel de destaque pela inovação e customização que promovem. Estas técnicas representam um fator diferenciador para a empresa, pois procuram desenhar soluções e experiências personalizadas, desenvolvendo produtos e/ou serviços com base nas necessidades e preferências do consumidor, com um potencial impacto na sua satisfação global. Dentro destes esforços, a geração e gestão de *leads* tem ganho relevância e é atualmente o foco principal das estratégias de aquisição de clientes de muitas empresas.

O trabalho de investigação desenvolvido sugere a aplicação de técnicas de *data mining* na otimização dos processos de gestão de *leads*, desde a captura, ao tratamento e conversão, com o objetivo de melhorar a eficácia da conversão final em clientes. Na literatura observou-se uma intensa aplicação destas metodologias em casos práticos de gestão de clientes, contudo, identificou-se uma lacuna relativamente a esta mesma aplicação em dados de *leads.* Nesse sentido, propôs-se uma metodologia que procura melhorar a eficiência das diferentes fases da gestão de *leads* (capturar, enriquecer, nutrir, rastrear, avaliar e converter), assim como suportar a tomada de decisão na segmentação de *leads* em critérios como: o número de *leads* a contactar, a seleção de *leads* com maior propensão e respetiva alocação nos canais de vendas.

Atendendo a vários estudos, as estatísticas demonstram que a maior dificuldade das empresas encontra-se precisamente nos processos de enriquecimento e nutrição da gestão de *leads*, refletindo-se na fraca qualidade dos dados e baixas taxas de conversão. Desse modo, a metodologia foca essencialmente no desenvolvimento de um modelo de propensão, que deve incluir um modelo de limpeza na fase de pré-processamento dos dados, e um modelo preditivo na fase de construção do modelo de classificação propriamente dito.

Assim, foi desenvolvido um caso de estudo na indústria das telecomunicações, onde foram aplicadas as metodologias propostas. Posteriormente, foram identificadas oportunidades de melhoria e sugeridas as respetivas soluções a curto e médio-longo prazo. Por fim, foram apresentados e discutidos os resultados em profundidade.

Palavras-chave: Gestão de *Leads; Data Mining;* Modelo Preditivo, Eficiência; Eficácia

Abstract

The growing market competitiveness has been pressing companies to improve the efficiency and effectiveness of their customer relationship management strategies. Specifically for customer acquisition, the direct marketing approaches have been attaining a prominent role due to their promotion of innovation and customization. These techniques represent a differentiating factor to the company since they seek to design personalized solutions and experiences, developing products and/or services based on the consumer's needs and preferences, which can lead to a potential impact on their global satisfaction. Within these efforts, the generation and management of leads have been gaining relevance and is currently the focus for customer acquisition strategies of many companies.

The developed investigation work suggests the application of data mining techniques in the optimization of leads management processes, from capture to conversion, with the objective of improving customer conversion effectiveness. In literature, it was observed an intense application of these methodologies in business cases regarding customer relationship management, however, it was identified a gap related to the application of these techniques in leads data. In this sense, it was proposed a methodology that aims to improve efficiency on the different stages of leads management (capture, enhance, nurture, tracking, assess and convert), as well as support decision-making regarding the segmentation of leads in criteria such as: number of leads to contact, selection of higher propensity leads and respective allocation to sales channels.

Considering several studies, statistics demonstrate that the biggest difficulties companies have are, precisely, the enrichment and nurture stages of leads management, reflecting on the lack of data quality and low conversion rates. Thereby, the methodology focuses essentially on the development of a propensity model, which includes a cleansing model in the data pre-processing phase, and a predictive model when building the classification model itself.

Therefore, a business case in the telecommunications industry was developed, in which the proposed methodologies were applied. Afterwards, several opportunities for improvement were identified and the respective short and long term solutions were suggested. Lastly, the results were presented and discussed in depth.

Keywords: Lead Management; Data Mining; Predictive Model; Efficiency; Effectiveness

Índice de Conteúdos

Índice de Figuras

Índice de Tabelas

Lista de siglas

AUC – *Area Under the Curve*

CED – *Contract End Date*

CRISP-DM - *Cross Industry Standard Process for Data Mining*

CRM – *Customer Relationship Management*

KDD – *Knowledge Discovery in Databases*

MLP - *Multi-layer perceptron*

PC – *Principal Components*

PCA – *Principal Components Analysis*

ROC - *Receiver Operating Characteristic Curve*

SEMMA – *Sample, Explore, Modify, Model, Assess*

1. Introdução

1.1. Motivação e Questão de Investigação

Nas últimas décadas, a competição no sector das telecomunicações tem-se tornado cada vez mais feroz e desafiante. Alguns autores consideram que esta é uma consequência de uma característica particular deste sector: o facto de não haver perspetivas de crescimento de mercado por existir uma restrição geográfica associada. Como resultado, a vantagem competitiva das operadoras de telecomunicações pode advir de duas fontes, ou do ganho de quota de mercado, ou da exploração de novos mercados através da disrupção de produtos e serviços. A pressão exercida pela forte concorrência fez com que as empresas procurassem sempre intervir de forma rápida, eficaz e eficiente, mantendo o foco na qualidade do serviço e na satisfação do cliente (Zhang, Yang, Shi, & Lu, 2008). Principalmente na era em que o preço já não representa um fator diferenciador para as empresas do sector (Guifang & Youshi, 2010), surgem na literatura vários artigos que visam o estudo do comportamento e preferências dos clientes (Wang, Sanguansintukul, & Lursinsap, 2008). Com o objetivo de melhorar a gestão de relação com o cliente e inovar pela personalização da oferta e da experiência, estes estudos incidem essencialmente no desenho de soluções customizadas (Gupta, Wasid, & Ali, 2016), procurando suportar as estratégias das empresas.

A procura de soluções inovadoras, alinhadas com as tendências de mercado, resultou na idealização e adoção de novas estratégias de aquisição de clientes. Surgiu assim o conceito de captura de *leads* como estratégia para aquisição de novos clientes (Goldie, 2007; Krozel, 2019). Considera-se uma *lead* um conjunto de informação útil, resultante da expressão de interesse num produto ou serviço por parte de um indivíduo (Willis & Flo, 2016). Quando este indivíduo demonstra potencial interesse em tornar-se cliente, são recolhidos dados que permitam contactá-lo posteriormente, e assim explorar todas as potenciais oportunidades de negócio (Willis & Flo, 2016; Bairstow, 2016). Os processos de criação e captura de *leads* têm ganho bastante visibilidade nas empresas, principalmente no âmbito do marketing digital. Porém, há vários canais de marketing responsáveis pela geração de *leads*, desde os canais *offline*, como a televisão, revistas e o telemarketing, aos canais *online* que fazem uso das tecnologias digitais como redes sociais, vídeos, blogs, *newsletters*, entre outros.

Na literatura, vários investigadores abordam a gestão de *leads c*omo uma estratégia inovadora no âmbito da aquisição de novos clientes. Para muitos, esta é uma forma de rentabilização dos canais digitais, possibilitando o contacto com uma base estável de possíveis clientes (Krozel, 2019). No entanto, estudos

indicam que ainda não existe uma consciencialização generalizada de como tirar o máximo partido da informação disponível. De facto, já em 2007 a maioria das empresas (82%) reconhecia que a captura de *leads* era uma área em crescimento, porém, apenas 44% afirmava estar a explorar eficazmente o valor das *leads* com o objetivo de crescimento e consolidação do negócio (Goldie, 2007). Segundo consta num relatório mais atual, apresentado pela empresa Velocify, 70% das *leads* geradas nunca chegam a receber um acompanhamento apropriado, e mais de 50% nunca chega a ter uma segunda interação com a empresa (Velocify, 2012). Estas estatísticas revelam que as empresas ainda são bastante ineficientes nos processos de gestão e conversão de *leads*.

Ainda neste âmbito, um estudo conduzido pela consultora *Ernst & Young* demonstrou que 58% das empresas estão descontentes com a eficácia dos processos, enfatizando como principais causas a fraca qualidade dos dados produzidos e das ferramentas digitais utilizadas (Silverstein, 2012). Em 2019, Krozel expõe outros casos nos quais as taxas de conversão são praticamente nulas devido, também, ao défice de qualidade das *leads*. Nesse sentido, Coe (2004b) e Blattberg et. al., (2008) estudam o impacto da qualidade das *leads* nos resultados obtidos e demonstram que o grau de qualificação é inversamente proporcional à taxa de sucesso esperada. Os autores apontam também que taxas de conversão baixas fazem aumentar substancialmente o custo de aquisição por cliente. Assumindo que é impossível para as empresas aumentar o volume de chamadas realizadas, pois este canais já funcionam perto da capacidade máxima (Coe, 2004b), a única alternativa é melhorar a eficácia e a qualidade das chamadas. Assim, com o objetivo de angariar o máximo número de clientes, minimizando o investimento em campanhas de marketing, surge nas empresas uma maior preocupação com a qualificação de *leads*, com a eficiência dos processos e com o aumento das taxas de conversão.

Embora o objetivo de aumentar a eficácia de campanhas, reduzindo o número de contactos, seja comum a empresas de diversos sectores (Shaoling & Yan, 2008), Compton (2012) afirma que cada empresa deve adotar uma abordagem customizada e adaptada às suas necessidades. Isto significa que não existe uma solução ou modelo ótimo que se aplique a todos os cenários, sendo necessário compreender o funcionamento e as especificidades de cada negócio.

Contudo, no que diz respeito a tarefas frequentes e que podem ser automatizadas, existe um consenso geral de que técnicas de *data mining* podem facilitar e otimizar os processos em casos reais (Wang, Sanguansintukul, & Lursinsap, 2008; Yang, Zhang, & Zuo, 2008; Zhang, Yang, Shi, & Lu, 2008; Zhao, Wu, & Gao, 2008; Shaoling & Yan, 2008; Guifang & Youshi, 2010; Moro, Laureano, & Cortez, 2011). As investigações centram-se essencialmente no desenvolvimento de modelos preditivos no âmbito da gestão da relação com o cliente, suportando análises como a propensão para aquisição de um novo produto, aceitação de campanhas, aumento de valor ou abandono de clientes. Contundo, não há

referências na literatura da aplicação destas técnicas ao conceito de *leads e* de potenciais clientes. Idealmente, aplicados à gestão de *leads*, estes métodos poderiam ser utilizados para estimar a propensão de conversão de cada *lead*, e assim, gerir de forma mais eficiente os recursos e garantir uma maior taxa de sucesso.

Teoricamente, um modelo de propensão tem como objetivo classificar um novo registo de acordo com a probabilidade de ocorrência do evento de sucesso – que neste caso é a conversão. No seu artigo, Eggert & Serdaroglu (2011) comprovam que a aplicação de um modelo preditivo a um conjunto de dados de clientes, para identificar aqueles com maior probabilidade de conversão, aumenta a produtividade das atividades de telemarketing. Assim, a identificação do conjunto de *leads* com maior probabilidade de conversão permitiria, simultaneamente, reduzir o número de contactos efetuados a potenciais clientes e aumentar a probabilidade de conversão, rentabilizando o investimento de cada campanha.

Assim, o trabalho desenvolvido nesta dissertação sugere a aplicação de um modelo preditivo no processo de gestão de *leads.* A investigação foi conduzida no sentido de demonstrar a aplicabilidade de técnicas de *data mining* neste tipo de problemas, produzindo um modelo de propensão e segmentação com base na informação escondida nos dados das *leads*. O objetivo final é proporcionar uma solução alinhada com as necessidades da empresa, que permita suportar o processo de decisão e torná-lo menos arbitrário. Deste modo, apontam-se como questões de investigação desta dissertação as seguintes:

1. Como é que as técnicas de *data mining* podem auxiliar a tomada de decisão no processo de segmentação de *leads*?
2. Como aumentar a taxa de conversão de *leads*?

1.2. Objetivos

Atendendo às questões de investigação, estabeleceram-se como principais objetivos desta dissertação os seguintes:

- Compreender o atual fluxo de processos na gestão de *leads*, desde a captura, ao tratamento e conversão;
- Identificar melhorias nos processos, a curto e médio-longo prazo, que visem a melhoria da qualidade dos dados;
- Desenvolver um modelo de limpeza, com o objetivo de melhorar a qualidade dos dados armazenados;

- Desenvolver um modelo de propensão, recorrendo à aplicação de técnicas de *data mining* a dados reais de uma empresa, simulando um processo de tomada de decisão na segmentação de *leads.*
- Analisar os resultados obtidos e propor um conjunto de soluções possíveis, demonstrando o potencial aumento da taxa de conversão esperada.

1.3. Metodologia

Considerando as questões de investigação e os objetivos definidos nesta dissertação, foi definida a metodologia de investigação apresentada na Figura 1.1.

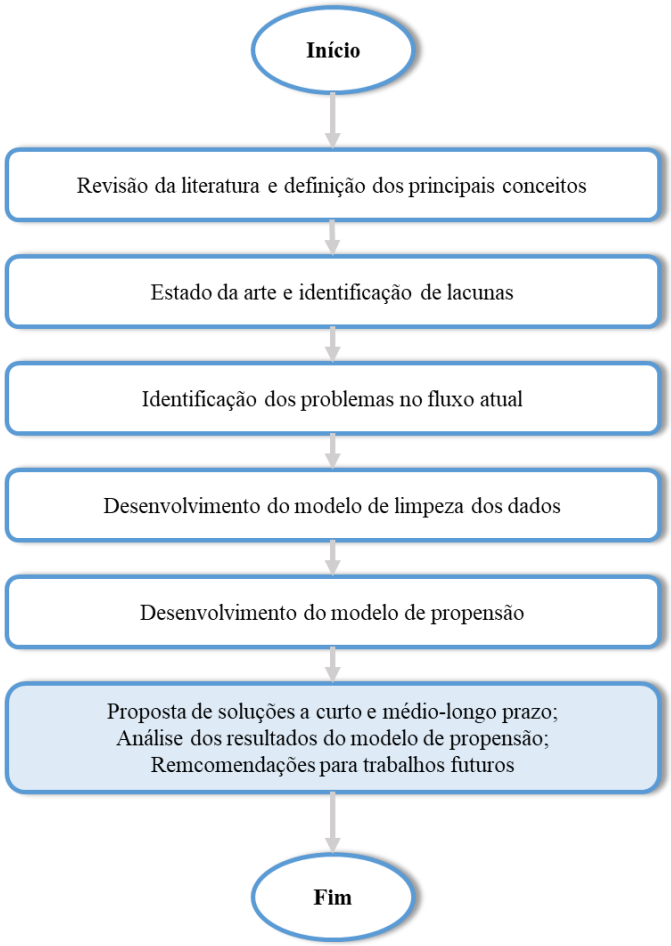

Figura 1.1 - Metodologia da Dissertação.

Conforme apresentado na figura 1.1, conduziu-se uma revisão da literatura existente nas áreas de Gestão da Relação com o Cliente (CRM), gestão de *leads* e *Data Mining*. O estudo do CRM enquadra-se nas atividades de marketing direto para a aquisição de clientes e no aparecimento de novas estratégias como a gestão de *leads.* Sendo este o foco da dissertação, houve necessidade de aprofundar a investigação acerca dos modelos de gestão de *leads* propostos na literatura, identificando as principais vantagens e problemas. A revisão das técnicas de *data mining* surge na sequência da proposta dos vários autores para a aplicação destas metodologias na resolução de problemas de classificação em casos reais.

De seguida, procurou-se realizar um estado da arte relativamente à aplicação de *data mining* em problemas do âmbito da gestão da relação com os clientes, com o objetivo de melhorar a eficácia dos contactos. Esta etapa permitiu ainda identificar uma lacuna na literatura, relativamente à aplicação destas técnicas em dados de *leads,* sendo por isso o tema endereçado na questão de investigação.

A terceira etapa compreendeu a análise da implementação da solução atual e respetiva identificação de problemas ao nível dos processos. Os restantes passos representam a contribuição da presente dissertação para o caso em estudo. Primeiramente, a contribuição inclui a proposta de soluções, a nível dos processos e da implementação técnica, para os problemas previamente identificados. De seguida, o desenvolvimento de um modelo de limpeza de dados garantiu a melhoria significativa da qualidade dos dados armazenados, potenciando, assim, os resultados esperados pelo modelo de propensão.

Por fim, foram apresentados e interpretados os resultados do modelo de propensão, com o objetivo de auxiliar o processo de segmentação de *leads,* e sugeridas recomendações para trabalhos futuros.

1.4. Estrutura da Dissertação

Esta dissertação encontra-se organizada em 6 capítulos. O primeiro capítulo é uma introdução ao trabalho desenvolvido, descrevendo a motivação, os objetivos, a metodologia seguida e a estrutura do documento. Os capítulos 2, 3 e 4 constituem a revisão bibliográfica da dissertação, incidindo nas áreas de Gestão da Relação com o Cliente (CRM), Gestão de *Leads* e *Data mining.*

O capítulo 2 dá uma visão geral do conceito de CRM e das diferentes abordagens, evidenciando a definição de alguns conceitos chave para a compreensão do trabalho. No capítulo 3 fez-se uma introdução ao mercado das telecomunicações e ao seu foco em estratégias de aquisição de clientes. Consequentemente, analisou-se o conceito de *leads,* o seu modelo de funcionamento e foram identificados os principais fatores que contribuem para a conversão de *leads* em clientes. O estudo das

técnicas de *data mining* foi endereçado no capítulo 4. Este capítulo fornece fundamentação teórica às metodologias aplicadas na formulação de propostas e na construção do modelo preditivo.

No capítulo 5 foi analisado o caso de estudo e identificados os principais problemas a nível processual e técnico. A deteção de problemas conduziu ao desenvolvimento do modelo de limpeza e de propensão, também descritos neste capítulo.

Por fim, no capítulo 6, foram apresentadas as respetivas soluções para os problemas identificados, e discutidos os resultados do modelo de propensão. Com o objetivo de suportar a tomada de decisão no processo de segmentação de *leads,* foram formuladas as possíveis soluções finais. Finalmente, refletiu-se acerca das limitações do caso em estudo e teceram-se recomendações para investigações futuras.

2. Gestão da Relação com o Cliente

A expressão Gestão da Relação com o Cliente, do inglês *Customer Relationship Management* (CRM), é frequentemente utilizada na literatura de marketing contemporânea. Apesar de surgir no início da década de 90, não foi ainda obtido um consenso relativamente à sua definição entre os vários autores (Buttle, 2009a; D'Haen & Poel, 2013; Guo & Qin, 2017; Guifang & Youshi, 2010; Yang, Zhang, & Zuo, 2008; Wang, Sanguansintukul, & Lursinsap, 2008). Contudo, a maioria dos conceitos apresentados têm em comum algumas características, que vão de encontro com a definição sugerida em 1990, pelo Grupo Gartner, definindo o CRM como uma estratégia empresarial que visa a otimização da receita e do lucro, enquanto promove a satisfação e lealdade do cliente (Guo & Qin, 2017; Guifang & Youshi, 2010).

Desde então, o conceito tem evoluido em várias vertentes. Guo & Qin (2017) acrescentam que esta estratégia centrada no cliente passa também por recolher, analisar e explorar informação compreensiva das interações entre ambas as partes, de modo a racionalizar a alocação de recurso de acordo com a necessidade, procura e o próprio comportamento do cliente. Deste modo, as empresas desenvolvem as suas capacidades de comunicação com o cliente, melhorando a sua experiência e a sua afinidade com marca (Guifang & Youshi, 2010). Outros autores focam a componente comercial de aquisição e retenção de clientes, assim como a maximização do valor de cliente a longo prazo (D'Haen & Poel, 2013). Os processos de CRM permitem a gestão e coordenação das interações dos clientes através dos diferentes canais comerciais, departamentos e linhas de negócio, ajudando as organizações a maximizar o valor de cada interação, conduzindo a um desempenho superior (Buttle, 2009a; D'Haen & Poel, 2013).

Na perspectiva de valor de cliente, muitas empresas já reconhecem que nem todos os clientes são igualmente rentáveis, assim como nem todos têm o mesmo grau de afinidade (Yang, Zhang, & Zuo, 2008). Na opinião dos autores Wang, Sanguansintukul e Lursinsap (2008), as técnicas de CRM devem, portanto, assistir e suportar análises de propensão para aquisição, aumento de valor ou abandono, garantindo a eficácia da comunicação e a eficiência na utilização de recursos.

Tipicamente as técnicas de CRM adotam um papel central nas estratégias de aquisição e retenção de clientes (Yang, Zhang, & Zuo, 2008), sendo que, por vezes, as estratégias de aquisição são negligenciadas na literatura, face às estratégias de retenção, por representarem um esforço e um custo maior para as empresas (Buttle, 2009b ; Berry, 2002). Na literatura são referidos dois tipos de marketing, o Marketing Relacional e o Marketing Direto, que acabam por estar diretamente relacionados com as técnicas de retenção e de aquisição, respetivamente.

O conceito de Marketing Relacional aparece pela primeira vez na literatura, em 1983, com Leonard Berry. Até à data, a grande preocupação das unidades de marketing das empresas era atrair novos clientes, sendo mínimos os esforços e os recursos aplicados para reter clientes existentes. Contudo, no seu artigo, o autor defende que as empresas beneficiam, tanto ou mais, ao manter clientes, assim como ao atrair novos. Como refere o autor, a carteira de clientes pode ser vista como um fluxo de entradas e saídas, portanto, independentemente do número de novos clientes angariados, é importante que as saídas sejam mínimas, mantendo um balanço positivo. Isto consegue-se promovendo a confiança e a lealdade dos clientes com a marca. (Berry, 2002). Assim, na base das atividades do Marketing relacional integram-se também os objetivos dos CRM, nomeadamente o cultivo e o enriquecimento da relação com o cliente através da prestação de um serviço de qualidade, personalizado, diferenciador e na proposta de valor acrescentado. Segundo Berry (2002), este investimento justifica-se sempre que houver um desejo explícito do cliente para manter um serviço, quando a seleção do fornecedor é controlada pelo cliente, e sempre que houver fornecedores alternativos.

No entanto, e apesar da importância de uma forte retenção, a aquisição de novos clientes deve ser também um foco para as empresas, como forma de compensar clientes perdidos, ou no âmbito de exploração de novos mercados (D'Haen & Poel, 2013). Para tal, as empresas devem desenvolver, cada vez mais, os seus produtos e serviços com base nas necessidades e preferências do consumidor (Zhang, Yang, Shi, & Lu, 2008), pelo que, tem-se observado um investimento acrescido na análise comportamental do mercado como suporte à decisão. Guifang & Youshi (2010) referem que no sector das telecomunicações, por exemplo, tem sido feita uma aposta na aquisição de dados de clientes, com o objetivo de cobrir os padrões comportamentais nos vários âmbitos. Contudo, quando maior o volume de dados recolhidos, mais complexo se torna o processo de tratamento e exploração dos dados, emergindo a necessidade de tecnologia mais avançada e novas ferramentas (Guo & Qin, 2017). No âmbito do Marketing Direto há vários trabalhos que recorrem a técnicas de *data mining* para obter melhores resultados na aquisição e conversão de novos clientes (Ling & Li, 1998; Hu, 2005; Lin, Wan, & Pu, 2010).

Uma vez abordado o conceito e o papel do CRM nas empresas, integrado nas estratégias de marketing relacional e direto, nos subcapítulos seguintes debate-se com maior detalhe as diferentes abordagens do CRM e os processos de aquisição de novos clientes, visto serem os temas mais relevantes como resposta aos objetivos definidos inicialmente.

2.1. Diferentes abordagens de CRM

A gestão de relação com o cliente foi definida como um sistema de informação integrado, que é utilizado para planear, programar e controlar as atividades comerciais. A sua ações abrangem todos os aspetos que envolvem lidar com clientes e potenciais clientes (Guifang & Youshi, 2010). O principal objetivo do CRM é proporcionar um crescimento sustentado a longo prazo e aumentar a rentabilidade das ações comerciais através da compreensão do comportamento dos clientes (Buttle, 2009a).

Nesta perspetiva, em que o CRM desempenha o papel de ponte integradora entre os sistemas de informação e a complexidade dos requisitos de negócio, são sugeridas na literatura 3 âmbitos de aplicação do CRM: estratégico, operacional, analítico (Buttle, 2009b; Guifang & Youshi, 2010). Conforme definido por Buttle (2009b), o CRM estratégico tem como pilares as estratégias empresariais centradas no cliente, com o objetivo de atrair e manter os clientes mais valiosos. Nesta abordagem, a cultura de valores centrados no cliente tem origem nos cargos de liderança e reflete-se em medidas como a alocação de recursos em atividade que aumentam o valor do cliente, ou o reconhecimento e recompensação de comportamentos que promovem a satisfação e a retenção de clientes. O CRM operacional foca na automação de processos, como é o exemplo das vendas, marketing, atendimento e suporte ao cliente. Por sua vez, as técnicas de CRM analítico servem os processos de transformação e tratamento de dados de clientes que entregam às respetivas áreas os *insights* e o conhecimento acionável para fins estratégicos. Recorrendo a metodologias como as de *data mining*, esta vertente procura a facilitação de processos, como é o caso dos processos de aquisição de clientes (Buttle, 2009b).

Atendendo à questão de investigação levantada, é do interesse do estudo aprofundar a análise das competências, vantagens e oportunidades que advêm da aplicação de técnicas de CRM analítico nas organizações. Segundo Buttle (2009a), é do âmbito do CRM analítico a captura, armazenamento, extração, integração, processamento, interpretação, distribuição, utilização e reporte dos dados relacionados com os clientes, com o objetivo de valorizar o cliente e a empresa. Os pilares do CRM analítico assentam em toda a informação armazenada nos repositórios da empresa, resultante da relação estabelecida entre o cliente e a empresa, como o histórico de compras, os dados transacionais e financeiros, a resposta a campanhas de marketing, os programas de fidelização, as ações de suporte, entre outras. Dado o volume, a complexidade e a heterogeneidade dos dados, as empresas têm que recorrer à aplicação de técnicas de *data mining* para extrair conhecimento válido e útil, que permita decidir, por exemplo qual a abordagem de vendas a adotar para cada grupo de clientes.

Do ponto de vista de cliente, o CRM analítico tem potencial para oferecer soluções customizadas, no tempo certo, promovendo a satisfação e a lealdade do cliente. Do ponto de vista da empresa, este oferece programas de retenção e aquisição mais eficazes e poderosos.

2.2. O papel do Marketing Direto na Aquisição de Clientes

Como afirma Greenyer (2000), as grandes empresas têm agora a necessidade de interagir com o cliente da forma mais tradicional. Tipicamente em pequenos negócios, o vendedor sabia exatamente o nome de todos os seus clientes, o que eles gostavam ou não gostavam, o que consumiam regularmente, e eram capazes de recomendar os produtos certos com base na observação do seu comportamento (Greenyer, 2000). É este nível de personalização de oferta e de conhecimento do consumidor que as empresas procuram, e que tem produzido efeitos na aplicação tecnológica nas áreas de marketing.

Anteriormente, quando não existia uma consciencialização para a alocação eficiente de orçamento e recursos nos processos de marketing, a abordagem das empresas focava-se na promoção massiva e não direcionada de produtos e serviços O marketing massivo utilizava meios de media mais abrangentes como televisão, radio, revistas e outro tipo de publicidade online sem público alvo definido (Ling & Li, 1998). Porém, em termos de aquisição, impactar todos consumidores com a mesma oferta não é uma estratégia eficaz nem eficiente (Bhattacharyya, 1998). Na realidade atual, em que os produtos são cada vez mais atraentes e o mercado é tão competitivo, estratégias de marketing indiferenciado não surtem o efeito desejado (Hu, 2005).

Por sua vez, o marketing direto identifica um público alvo que à partida tem necessidade de um produto ou serviço, já demonstrou interesse em adquirir (Ling & Li, 1998), ou reúne um conjunto de características similares a atuais utilizadores (Moro, Laureano, & Cortez, 2011). Na literatura, são vários os autores que referem a aplicação de modelos de *data mining* como um contributo na identificação e seleção de clientes para campanhas de marketing específicas (Hu, 2005; Bhattacharyya, 1998; Ling & Li, 1998; Lin, Wan, & Pu, 2010). Se o modelo for bem definido e adequado às necessidades de negócio, a empresa poderá contactar um grupo menor de pessoas, mas com maior potencial para aceitar a oferta (Bhattacharyya, 1998). De um modo geral estes estudos tomam uma abordagem de classificação, onde o objetivo é construir um modelo preditivo que permita prever o resultado para cada registo, dentro de um conjunto de resultados pré-definidos (Moro, Laureano, & Cortez, 2011).

Atualmente, as empresas têm a capacidade de armazenar os dados relativos à caracterização pessoal e comportamental dos seus clientes, através das interações nos diversos canais, tendo ao seu dispor todo

o conhecimento necessário para personalizar a oferta e comunica-la ao cliente através do canal mais indicado, seja email, telemarketing ou *online.* Quando analisados corretamente, esses dados transformam-se no mesmo conhecimento utilizado no atendimento tradicional, mas numa escala muito maior (Greenyer, 2000).

A adoção de novos canais surge da necessidade de explorar novos mercados, mas também da necessidade de comunicar com clientes da forma mais eficaz e com menor custo possível (Chen, Kou, & Shang, 2014). A utilização do meio *online,* por exemplo, passou a ter um peso significativo na recolha de dados, no contacto com o cliente e na promoção comercial (Chen, Kou, & Shang, 2014). Numa estratégia multicanal é fundamental compreender o funcionamento de cada canal individualmente e as respetivas interações, de modo a identificar possíveis intersecções ou cruzamentos na atividade, que resultem em conflitos de informação. Essencialmente, é importante analisar a estrutura como um todo e avaliar o desempenho global. Para além disso, a identificação de forças e fraqueza permite a aplicação de ações de melhoria contínua nos processos (Kim, 2007). Porém, a integração dos diversos fluxos de informação pode ser uma tarefa complexa, visto que cada processo produz dados com qualidade, formato e cadência distintas. Alguns dos desafios da gestão multicanal residem, então, na definição de prioridades, na criação de sinergias, na identificação de um modo de cooperação ótimo face ao modelo de comissionamento de cada canal e na integração técnica dos vários sistemas de informação (Chen, Kou, & Shang, 2014).

A introdução dos canais *online* na estratégia das empresas, veio agitar a forma como os dados são capturados, introduzindo maior heterogeneidade no formato e qualidade dos dados. Para muitas, este canal é o principal ponto de vendas, enquanto que para outras é visto essencialmente como um ponto de atração, entretenimento e angariação de contactos para possíveis conversões em canais *offline* (Krozel, 2019). No entanto, independentemente de o meio de aquisição ser *online* ou *offline*, o processo de aquisição decorre na sequência de várias fases, que é definido por vários autores como o Funil de Vendas ou Funil de Conversão (Patterson, 2007; Yu & Cai, 2007; Coe, 2004a). Nos estudos, são apresentadas várias conceptualizações deste processo, sendo utilizadas terminologias e definições diferentes para cada uma das fases. No entanto, as principais divergências ocorrem, não ao nível estrutural, mas na definição dos conceitos de *lead* e *prospect*: alguns autores colocam a fase de *prospect* antes da fase de *lead* (Coe, 2004a; Metzger, 2005), enquanto outros colocam a fase de *leads* antes da fase de *prospect* (Patterson, 2007; Gillin & Schwartzman, 2011).

Dada a investigação realizada, optou-se e prosseguiu-se com a estrutura e definições apresentadas na figura 2.1, por ser a mais relevante e alinhada com os procedimentos praticados na empresa.

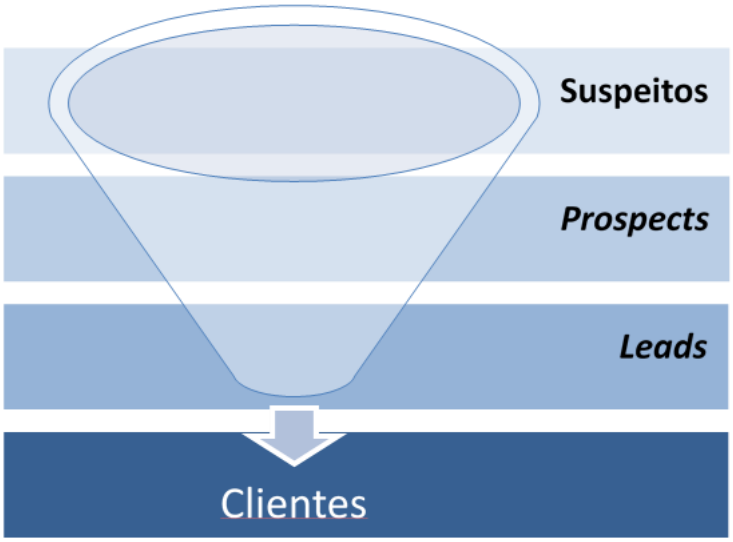

Figura 2.1 - Estrutura do processo de aquisição de clientes: Funil de Conversões. Adaptado de D'Haen & Poel (2013).

Conforme exibido na figura 2.1, os "suspeitos" representam todos os potenciais clientes disponíveis que, embora não tenham demonstrado interesse, são elegíveis para a aquisição de determinado produto ou serviço (Coe, 2004a). A transformação de "suspeito" em "*prospect"* ocorre quando o indivíduo atinge certas características pré-definidas, como resposta a uma campanha de marketing, por exemplo. A distinção entre "*prospect" e "lead"* reside no nível de qualificação. Uma *lead* ocorre quando um indivíduo demonstra um interesse explícito no tal produto ou serviço, e são reunidas as informações necessárias que permitem assegurar o nível de qualificação (Coe, 2004a).

Embora estes sejam conceitos novos em muitas empresas, verifica-se que atualmente, as estratégias de aquisição de clientes focam-se essencialmente na conversão de *leads* em clientes reais, descurando a fase de enriquecimento e qualificação da informação. Os estudos indicam que taxas de conversão baixas aumentam o custo de aquisição de clientes (Blattberg, Kim, Kim, & Neslin, 2008) e que a ineficácia recorrente da tomada de decisão nestes processos reduz o valor global de uma empresa ao longo do tempo (D'Haen & Poel, 2013). Assim, aumentar a taxa de conversão é também reduzir o custo por aquisição (Coe, 2004b).

3. O papel das *Leads* no sector das Telecomunicações

O sector das Telecomunicações é caracterizado pela elevada competição, pela orientação para o cliente e pelo enorme volume de dados produzidos a cada minuto (Zhao, Wu, & Gao, 2008; Zhang, Yang, Shi, & Lu, 2008; Guifang & Youshi, 2010). De facto, a orientação para o cliente acaba por ser uma consequência dos outros dois fatores: dada a forte competição do mercado e o desafio contante de tornar as operações mais eficientes, a indústria das telecomunicações foi uma das pioneiras na utilização dos dados como forma de valorização dos seus produtos e serviços (Guifang & Youshi, 2010). A aplicação de técnicas de *data mining* às áreas de marketing e CRM tem sido um assunto abordado e estudado por vários autores (Zhao, Zhang, Wang, Zhang, & Wang, 2014; Zhao, Wu, & Gao, 2008; Zhang, Yang, Shi, & Lu, 2008; Wang, Sanguansintukul, & Lursinsap, 2008; Lu, Lin, Lu, & Zhang, 2014).

Porém, o mercado das telecomunicações tem a particularidade de ser restrito a uma área geográfica, e, portanto, não ser um mercado em expansão. Como consequência, sempre procuraram marcar a diferença na disrupção dos produtos e na qualidade dos serviços. Pela forte pressão de crescer em quota de mercado, houve também um grande investimento e especialização nas estratégias de aquisição de clientes.

Vários autores referem a gestão de *leads* como uma estratégia revolucionária para muitas empresas e inovadora no âmbito da aquisição de novos clientes (Goldie, 2007; Krozel, 2019). Para muitos esta é uma forma de rentabilização dos canais digitais, proporcionando contactos com uma base estável de possíveis clientes (Krozel, 2019). No entanto, estudos indicam que ainda não há uma consciencialização generalizada de como tratar e extrair o máximo valor das *leads* produzidas. De facto, ainda em 2007, 82% das empresas reconheciam que a captura de *leads* era uma área em crescimento, porém, apenas 44% afirmava estar a explorar eficazmente o valor das *leads* com o objetivo de crescimento e consolidação do negócio (Goldie, 2007). Citando Monat (2011, p. 192), o autor considera que as *leads* são a força vital das empresas, no entanto, ainda se baseiam muitas vezes em conjunturas ou intuições para determinar quais as *leads* que têm maior probabilidade de converter em novos clientes. No seu trabalho, o autor sugere um modelo quantitativo, baseado na informação inerente e capturada nas próprias *leads*, que permita prever se as *leads* irão converter.

Neste capítulo são apresentadas as circunstâncias e especificidades do setor das telecomunicações, como formar de examinar as potencialidades e possíveis vantagens de aplicação de um modelo de gestão de *leads.* De seguida foi abordada a evolução do conceito de *leads* e o papel que estas desempenham atualmente. Por fim, foi descrito o modelo teórico de gestão e conversão de *leads*, e identificados os principais desafios e problemas que as empresas enfrentam na adoção e aplicação destes métodos.

3.1. Introdução ao Mercado das Telecomunicações

Com a globalização da economia mundial e a internacionalização dos mercados, a competição na industria das telecomunicações tem-se tornado cada vez mais feroz e os desafios cada vez maiores (Guifang & Youshi, 2010). Para manter a vantagem competitiva, as operadoras procuram intervir de forma rápida, eficaz e eficiente, com foco na qualidade dos serviço e na satisfação do cliente (Zhang, Yang, Shi, & Lu, 2008). De facto, são vários os autores que referem estes como os três aspetos que caracterizam a realidade atual do sector das telecomunicações: concorrência elevada, orientação para o cliente e volume de dados (Zhao, Wu, & Gao, 2008; Guifang & Youshi, 2010).

Este sector tem ainda a particularidade de não ser um mercado em expansão, isto é, apesar de ter o poder de crescer em termos de quota de mercado e de se lançar em mercados diferentes através da disrupção de produtos e serviços, as empresas de telecomunicações operam num espaço geográfico limitado, onde a grande maioria dos habitantes já é consumidor de pelo menos um serviço fixo ou móvel. Esta particularidade implica que a aquisição de novos clientes resulte do cancelamento dos serviços com uma empresa concorrente. Mesmo em termos dos serviços oferecidos, as complexidades dos processos de aquisição têm algumas diferenças. Enquanto que a adesão de um serviço móvel é relativamente simples, no segmento fixo o processo é bem mais moroso e complexo. A adesão a um serviço fixo requer um conjunto de requisitos favoráveis, e envolve um processo de instalação e ainda a fidelização do cliente por períodos longos, o que também cria barreiras à portabilidade entre empresas.

Dado este cenário, as empresas do sector têm apostado na personalização das ofertas e no desenho de soluções à medida das necessidades e preferências dos clientes (Gupta, Wasid, & Ali, 2016), resultado do alinhamento entre as estratégias de CRM e *data mining*. Zhao, et. al., (2008) afirmam que a questão mais pertinente, dada a conjuntura apresentada, seria como adquirir mais conhecimento sobre os consumidores e como é que esse conhecimento iria conduzir as estratégias e as decisões da empresa. O estudo dos três fatores, e da forma como estes se complementam, pretende atender à questão levantada pelos autores.

Primeiramente, as operadoras detiveram um investimento tecnológico com o objetivo de expandir a capacidade de armazenamento e de preparar os sistemas para o volume de dados produzidos. No entanto, embora exista suporte tecnológico, muito do conhecimento útil para a área comercial, como as características de clientes e padrões de consumo, continuam escondidos no enorme volume de dados armazenado (Zhang, Yang, Shi, & Lu, 2008). Ou seja, o tratamento analítico dos dados continua a ser um dos grandes desafios para estas empresas. Guifang & Youshi (2010) identificam o aparecimento da internet como o grande impulsionador do crescimento exponencial de dados no sector *mobile*.

Como já foi referido, na literatura há vários autores que sugerem a aplicação de técnicas avançadas de *data mining* nesta área, comprovando bons resultados em casos reais (Wang, Sanguansintukul, & Lursinsap, 2008; Yang, Zhang, & Zuo, 2008; Zhao, Wu, & Gao, 2008; Shaoling & Yan, 2008). Como resposta aos desafios, surgem também novas oportunidades para que as empresas possam extrair o máximo valor dos seus dados e assim, obter vantagem face à concorrência (Zhang, Yang, Shi, & Lu, 2008). Em resposta à apertada concorrência, elevaram-se também várias iniciativas orientadas para a satisfação do cliente. Dada a conjuntura atual e a facilidade de acesso a informação *online*, os consumidores aparentam estar cada vez mais esclarecidos e exigentes, elevando as expectativas relativamente ao serviço prestado e à experiência em cada interação. Em muitos casos, o preço deixou de ser um fator diferenciador entre concorrentes (Guifang & Youshi, 2010). É, portanto, essencial investir em mecanismos de gestão de relação com o cliente, e criar critérios de avaliação e métricas específicas que permitam monitorizar o valor de cada cliente, e prever futuras flutuações (Wang, Sanguansintukul, & Lursinsap, 2008).

Resumidamente, a estratégia para enfrentar os desafios da nova era passam pela implementação de tecnologias e técnicas modernas de gestão de informação; pela reestruturação da tomada de decisão nas áreas de CRM e marketing, passando a incluir métodos mais científicos; e pela conceptualização de produtos e serviços disruptivos e customizados, que promovam a fidelização e a satisfação dos clientes (Moro, Laureano, & Cortez, 2011).

3.2. Leads

Uma *lead* é um conjunto de informação útil, resultante da expressão de interesse num produto ou serviço por parte de um indivíduo. Dada a demonstração de interesse, considera-se uma *lead* uma oportunidade de contacto com potencial para efetuar uma venda (Willis & Flo, 2016; Bairstow, 2016). Tipicamente, a informação recolhida numa *lead* inclui uma forma de contacto com o potencial cliente, para que seja possível "fechar o negócio".

Atualmente, a dinâmica das empresas tem evoluído no sentido de promover e estimular a captura de *leads* nos vários canais de comunicação com o cliente, consequência do reconhecimento das *leads* como um bem valioso, uma grande oportunidade para o negócio (Willis & Flo, 2016), e a principal fonte de aquisição de novos clientes (Bairstow, 2016). Em 2012, Silverstein expõe que tem existido uma mudança substancial na complexidade do processo de criação de *leads*, como resultado do crescimento do mundo *online* e da proliferação do marketing digital. Como refere também Bairstow (2016), com a explosão da *internet* nas últimas décadas, as ferramentas *online* tornaram-se a fonte primária de criação de *leads*. Basicamente, antes do aparecimento dos *social media* e da abrangência de conectividades, a

geração de *leads* era tão simples como o preenchimento de um formulário. Para Bairstow (2016), o foco do *website* de qualquer empresa deve incidir na produção de *leads* de elevada qualidade, com potencial para se transformarem em clientes de elevado valor.

A conjuntura atual é bastante mais complexa, integrando uma maior variedade de canais e de abordagens para a recolha de informação. Embora a introdução do mundo *online* dinamize o fluxo de aquisição, o objetivo final é conduzir os potenciais consumidores a tornarem-se clientes, o que pode tornar-se mais difícil devido à qualidade dos dados adquiridos. Frequentemente, os profissionais de marketing definem este processo como um "Funil de Conversão", como foi mencionado no capítulo 2. Aqui, a ideia de funil aplica-se por ser um processo mais abrangente no início e mais restrito e específico no final. Através de um conjunto de técnicas de qualificação, os *prospects* iniciais são conduzidos ao longo do funil até cumprirem os requisitos de informação que os classifiquem como *leads.* Mesmo entre o conjunto de *leads* podem ser definidos vários níveis de qualificação, sendo que aquelas com maior qualidade são convertidas em clientes com maior recorrência (Silverstein, 2012). Gordon (2018) afirma que só é extraído valor do funil de conversão quando a abordagem da empresa é centrada na qualidade das *leads*, e não quando a estratégia consiste em atingir todas as *leads* com a mesma mensagens, repetidamente.

No entanto, o processo de compra não é um processo linear, e como tal, a qualificação dos contactos, identificando aqueles que estão prontos para concretizar a compra, é uma estratégia muito mais eficiente e eficaz do que os meios tradicionais (Gillin & Schwartzman, 2011). Porém, a conversão também depende da competência e eficácia dos canais *offline* (Bairstow, 2016). Normalmente, é nesta fase do funil que há a intervenção dos canais de telemarketing ou de vendas diretas assistidas para "fechar o negócio". (Silverstein, 2012). A capacidade de resposta e a simplicidade dos processos são fatores que têm um forte impacto na conversão, e que acabam por ser um teste à qualidade do serviço (Bairstow, 2016). Assim, os processos de captura, acompanhamento e conversão de *leads* são vistos como atividades essenciais e promotoras de oportunidades de negócio (Willis & Flo, 2016). Na próxima secção são apresentadas as várias etapas do ciclo de vida de um *lead.*

3.2.1. Ciclo de Vida de uma *Lead*

De acordo com as referências apresentadas anteriormente, as *leads* são consideradas elementos vitais para o sucesso de qualquer organização (Silverstein, 2012). Aferiu-se também que, devido ao forte crescimento dos sistemas de informação e a das ferramentas *online*, os canais digitais tornaram-se as principais fontes de criação de *leads* e, consequente, de aquisição de clientes (Bairstow, 2016).

Como também já foi referido anteriormente, as *leads* não têm qualquer significado para o negócio se não houver uma conversão efetiva em clientes. No entanto, o processo de conversão não é simples e pode exigir várias interações com o cliente até se proporcionar o momento e as condições certas para a tomada de decisão (Gillin & Schwartzman, 2011). Na figura 3.1 foram resumidas as diferentes etapas desde o momento da captura da *lead* até ao momento da conversão (Gillin & Schwartzman, 2011; Silverstein, 2012; Samuels, 2013; Willis & Flo, 2016; Krozel, 2019). Embora cada autor apresente de forma diferente e atribua diferentes nomes a cada uma das etapas, há um consenso generalizado no que diz respeito aos 3 grandes grupos de ações: Quantidade, Qualidade e Conversão.

Figura 3.1 - Etapas do processo de conversão de um lead: Capturar, Enriquecer, Nutrir, Rastrear, Avaliar e Converter.

Tal como os nomes indicam, numa primeira fase o foco é a quantidade de *leads* geradas, seguida da fase de enriquecimento e melhoria da qualidade dos dados. Esta fase permite reunir a informação necessária que conduz à fase final de conversão em clientes. Observando a figura 3.1, pode inferir-se que a primeira fase consiste na captura de *leads*, enquanto que a segunda fase engloba as tarefas de enriquecer, nutrir e rastrear as *leads*. Como resultado, as *leads* ficam disponíveis para avaliação e conversão final. No entanto, note-se que o processo deverá ser completamente customizado e adaptado às necessidades e ao modelo de funcionamento de cada empresa, sendo que o processo apresentado na figura 3.1 funciona como mera representação da conceptualização teórica de cada fase (Compton, 2012). De seguida foram analisados os pressupostos e as características de cada etapa:

Capturar

Como refere Samuels (2013), de modo a aumentar o número de *leads* produzidas é importante identificar e quantificar as principais fontes de captura. A produção de *leads* pode acontecer de variadas maneiras, incluindo as formas mais tradicionais de publicidade, como televisão, rádio ou email, ou recorrendo aos canais digitais, como por via dos *social media*, *display advertising*, *paid search* ou outras fontes de *media* digital (Krozel, 2019). Silverstein (2012) define estes meios de comunicação como *outbound* e *inbound*, respetivamente, sendo que *outbound* são os meios que forçam uma mensagem aos utilizadores, enquanto que nos meios de *inbound* são os utilizadores que proactivamente procuram e interagem com

a informação disponibilizada. Estatísticas sugerem que os consumidores são mais atraídos por métodos *inbound*, resultando em maior taxa de eficácia para estes meios de captura de *leads* (Silverstein, 2012).

Gerar uma *lead* é o ato de identificar um possível cliente quando este executa uma ação concreta, expressando explicitamente o seu interesse num produto ou serviço (Silverstein, 2012). No mundo *online* esta demonstração de interesse pode adotar diversas formas, tais como um clique num anúncio, um *click to call,* um pedido de agendamento de chamada ou um preenchimento de formulário (Willis & Flo, 2016). Nesta fase, o essencial é recolher a informação básica que permita comunicar futuramente com a *lead,* seja o número de telefone, email ou morada. No entanto, o processo de captura de *leads* não se tem demonstrado simples nem económico, acabando por consumir muito tempo e recursos (Bairstow, 2016). Apesar de demonstrarem interesse, a maioria das *leads* não pretende avançar de imediato com a conversão. Como tal, é necessário investir na relação e cultivar o interesse do potencial cliente (Gillin & Schwartzman, 2011).

Enriquecer

Enriquecer os dados capturados, é a primeira etapa da fase que visa melhorar a qualidade da informação obtida acerca de cada *lead.* Enriquecer uma *lead* é adicionar informação qualificada sobre o consumidor, de forma a registar e categorizar o seu interesse. Esta informação minimiza o efeito de rejeição e torna mais eficiente o processo de aquisição (Krozel, 2019). O enriquecimento dos dados pode incluir características como o canal preferencial de contacto, estado civil, constituição do agregado familiar, formação académica, morada de residência, emprego, contratos atuais ou anteriores, ou qualquer outra informação que permita qualificar o interesse da pessoa e construir uma oferta à medida das suas necessidades. Quanto mais rica e valiosa for a informação coletada nesta etapa, mais fácil será o processo de nutrição e priorização subsequente dos contactos (Silverstein, 2012).

Nutrir

De facto, as atividades de enriquecimento e nutrição de *leads* podem ocorrer em simultâneo. Se uma *lead* for corretamente acompanhada, todas as interações são vistas como oportunidades para enriquecer a informação e afinar as próximas estratégias de contacto (Gillin & Schwartzman, 2011).

Nutrir uma *lead* envolve acompanhar os seus desenvolvimentos, mantendo-a ativa e interessada ao longo do tempo (Silverstein, 2012). Contudo, ao executar um programa de valorização de *leads* devem reconhecer-se as *leads* que são suficientemente importantes para o investimento de dinheiro e tempo, trabalhando no caminho de decisão do potencial cliente (Gordon, 2018).

Rastrear

Rastrear e acompanhar cada fase de uma *lead* acaba por ser uma tarefa transversal. Enquanto etapa da fase de qualificação, o objetivo é determinar a prioridade relativa baseada em critérios de negócio, para que seja desencadeado, no momento correto, as ações de conversão. Em termos tecnológicos é essencial a existência de plataformas que permitam medir e monitorizar as conversões, desde o início até à ultima fase do processo (Gordon, 2018), facilitando a gestão de *leads* e a identificação de potenciais oportunidades de negócio.

Avaliar

O processo de avaliação consiste numa junção de critérios: avaliar as *leads* produzidas e qualificadas, garantido que são selecionadas aquelas com maior propensão para tomar uma decisão final (Samuels, 2013); e alocar as *leads* selecionadas, com base num conjunto de regras de negócio, que definam o equilíbrio entre as oportunidades e os custos efetivos. Nesta fase, o foco reside em encontrar soluções de segmentação mais eficientes, que permitam alocar as *leads* corretas aos canais certos, no momento certo (Silverstein, 2012).

Converter

A conversão final é o ato de transformar uma *lead* capturada e qualificada num cliente. Este é o verdadeiro e principal objetivo de qualquer empresa. Compreender a forma como é realizada a venda, e os respetivos canais, é essencial para a definição de todo o processo de conversão (Samuels, 2013). Tipicamente, são utilizados 3 canais para efetuar as vendas finais, consoante as especificidades do negócio (Silverstein, 2012):

- *Online* – alguns negócios vendem principalmente ou exclusivamente *online.* Cada vez mais, as empresas apostam na componente *online* por ser o meio preferido dos clientes. Se assim for, o negócio tem alta dependência do marketing digital e as suas operações baseiam-se na assistência por chat *online,* respostas nas redes sociais, emails e questionários *online.*
- *Telemarketing* – utilização da capacidade de *call centers* para realizar campanhas de *outbound.*
- Vendas diretas – vendas assistidas por profissionais de vendas, em determinadas localizações como em casa, na empresa, entre outros.

Por fim, torna-se relevante calcular a taxa de conversão do processo, de forma a avaliar o sucesso das atividades levadas a cabo. Esta taxa representa o rácio entre o número de clientes que realmente realizou uma adesão, subscrição ou compra, e o número de inicial de potenciais clientes (Samuels, 2013).

Em suma, é importante perceber que, apesar de existir uma sequência lógica do processo de conversão, não há regras rígidas e rápidas que se apliquem a todas as empresas e a todas as *leads*. É por esse motivo que faz sentido adotar um sistema flexível, com capacidade de acomodar *leads* em vários estados do processo de conversão, e utilizar o programa de nutrição e valorização de *leads* para endereçar as necessidades dos possíveis clientes assim que forem surgindo (Silverstein, 2012). Assim, ressalta também a importância de ferramentas de personalização e customização de experiências, visto promoverem a satisfação do cliente e rentabilizarem o investimento global (Gordon, 2018).

3.3. Conversão de Leads – O impacto da qualidade dos dados

No capítulo anterior descreveram-se as várias etapas do processo de gestão de *leads*, que conduzem ao objetivo principal do negócio, a conversão e aquisição de novos clientes. No entanto, estudos demonstram que as empresas ainda são bastante ineficientes nos processos de gestão e utilização de *leads* (Silverstein, 2012). Num relatório apresentado pela empresa Velocify, foi possível apurar que cerca de 70% das *leads* geradas nunca chegam a receber um acompanhamento apropriado (Velocify, 2012) e que mais de 50% nunca chega a ter uma segunda interação com a empresa.

Num estudo realizado pela consultora *Ernst & Young* (Krozel, 2019), no âmbito dos processos de gestão de *leads* nas empresas, revelou-se que apenas 42% das empresas inquiridas estavam satisfeitas com a utilização das ferramentas digitais para a captura de *leads*. As principais razões indicadas para o descontentamento incidiam na fraca qualidade das ferramentas usadas e na fraca qualidade dos dados produzidos. Evidentemente que há vários fatores que contribuem negativamente para a qualidade da solução global e, consequentemente, têm um impacto negativo na eficácia dos processos (Silverstein, 2012). As observações efetuadas por Krozel (2019) expõem casos nos quais as conversões eram praticamente nulas devido à fraca qualidade das *leads*. Desde modo, pretendeu-se investigar quais os principais motivos assinalados na literatura para a degradação dos resultados do funil de conversão.

No que diz respeito à tecnologia aplicada nos processos de captura, tratamento e conversão de *leads*, o estudo realizado em 2018 pela EY identificou que em 55% dos casos as ferramentas são bastante antiquadas, lentas, com interfaces complexas e totalmente desatualizadas dos avanços e crescimentos desta área (Krozel, 2019). Na figura 3.2 apresenta-se uma adaptação do estudo realizado pela EY, onde se identificam os 5 principais motivos para o descontentamento com as ferramentas utilizadas na gestão de *leads*.

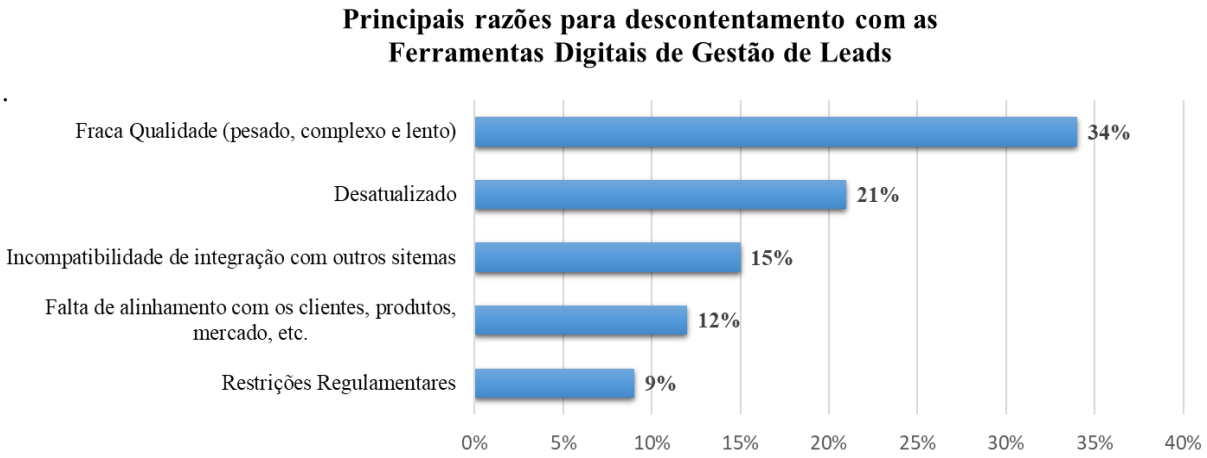

Figura 3.2 - Principais razões para o descontentamento com as ferramentas digitais de gestão de leads. Adaptado de Krozel (2019), LIMRA-EY Experienced Financial Advisor Study, 2018.

Pelos motivos mencionados na figura 3.2, é possível aferir que a desatualização das ferramentas tem repercussões a vários níveis: na compatibilidade com os restantes sistemas adotados pela empresa, dificultando as integrações com outros sistemas de informação; na capacidade de acompanhar a evolução do mercado, dos produtos e das necessidades dos clientes; a nível legal, na adaptação a novas normas e restrições regulamentares. Todos estes fatores influenciam o funcionamento normal dos processos requeridos, impedindo a obtenção dos resultados esperados (Krozel, 2019). Quando foi perguntado aos representantes de cada empresa presente no estudo, como classificaria a qualidade das *leads* produzidas nas suas empresas, apenas 30% atribuiu uma classificação positiva. Na figura 3.3 apresenta-se a distribuição das classificações atribuídas (Krozel, 2019).

Figura 3.3 - Classificação atribuída, pelos membros de cada empresa, à qualidade das leads produzidas. Adaptado de Krozel (2019), LIMRA-EY Experienced Financial Advisor Study, 2018.

Embora a percentagem de classificações baixas seja 28%, a maioria das empresas levantou vários motivos que justificam a frequência das elevadas taxas de insucesso. A primeira é a fraca qualidade, completude e veracidade dos dados. Muitas são as empresas que mencionam a heterogeneidade dos dados recolhidos e as falsas demonstração de interesse, que acabam por adicionar ruído aos dados. Por outro lado, mencionam vários problemas técnicos que conduzem, também, à produção de *leads* com fraca qualidade, como por exemplo a duplicação de registos e a falta de *standardização* dos campos de preenchimento (Krozel, 2019). Como Krozel (2019) aferiu, a maioria das empresas ainda não tinha procedimentos de nutrição e enriquecimento das *leads*, nem era instituída a preocupação de estimar a probabilidade de conversão de uma *lead* e assim criar audiências mais precisas e eficientes.

No artigo de 2008, Caller et al., refletem sobre a importância da qualidade dos dados nos processos de produção de *leads.* De facto, este tema tem ganho visibilidade perante as empresas pois estas começam a aperceber-se do impacto causado nos custos, no consumo de tempo e na perda de oportunidades de negócio (Olson J. E., 2003). Os autores Caller, Pallat e Darlow (2008) referem que as *leads* devem ser únicas para a organização, e como tal, devem ser planeadas e geridas de forma a preservar a qualidade e veracidade dos dados, evitando erros como a duplicação de registo e a introdução de informação errada. Embora seja difícil calcular os prejuízos associados à utilização de dados com má qualidade, Olson (2003) estima que os custos representem entre 15% e 25% do total do lucro. Por isso, é sugerida a utilização de alguns mecanismos de validação que permitam reduzir o número de inconsistências e dados incorretos (Caller, Pallat, & Darlow, 2008).

Com o intuito de encontrar possíveis soluções para os erros e inconsistências existentes nas bases de dados de *leads*, prosseguiu-se a investigação com o estudo dos fatores potenciadores de tais problemas. Identificaram-se os seguintes aspetos:

- Principalmente na recolha de *leads online*, não há a distinção entre uma *lead* criada por um indivíduo que está realmente interessado e fornece informação correta, e outra criada por um indivíduo que, por curiosidade, fornece dados errados e aleatórios (Silverstein, 2012).

- Numa empresa, a variedade de soluções para recolha de *leads* pode ser enorme. Quando não há um planeamento da arquitetura de informação em conformidade, a combinação dos dados pode ser complexa e a variabilidade dos dados existentes pode induzir a erros (Compton, 2012).

- Tradicionalmente, as áreas de marketing são responsáveis pela captura e qualificação das *leads*, enquanto que as áreas de venda garantem a sua conversão. Parte do problema ocorre quando o comissionamento da área de vendas baseia-se na quantidade de vendas ou respetivo valor

atingido. Quando esta situação ocorre, os vendedores não têm qualquer motivação ou compensação por qualificar as *leads*, nem por atualizar as áreas de marketing sobre os progressos alcançados no contacto. Como resultado, a organização perde controlo sobre as *leads* e sobre o funil de conversão em geral. Sem a informação vital que espelha o estado da *lead*, o marketing não é capaz de fechar o processo e analisar a eficácia do seu programa (Silverstein, 2012).

- A robustez e rápido crescimento de outras tecnologias fez crescer exponencialmente o volume de dados coletados, intensificando problemas como a variabilidade, duplicação, inconsistência e erros nos dados (Olson J. E., 2003).

- A propagação de dados incorretos por falta de sistemas de informação flexíveis, com mecanismos de correção simples, e adequados às necessidades de monitorização constante (Olson J. E., 2003).

- Falta de alinhamento entre as áreas de IT e Marketing na definição de requisitos e de critérios de aceitação durante o projeto de implementação. Embora o marketing tenha a responsabilidade de definir os requisitos funcionais das ferramentas, por vezes, a falta de conhecimento técnico resulta em falhas na especificação dos requisitos dos sistemas de captura e armazenamento dos dados. Olson (2003) afirma que estes erros poderiam ser minimizados se houvesse maior cooperação e envolvimento das equipas de IT no processo.

- Os sistemas utilizados pelas equipas de *telemarketing* nas campanhas *outbound* não estão preparados para integrar automaticamente o *feedback* resultante da campanha. Isto significa que não é adicionada nova informação às *leads*, não existindo atualização, enriquecimento ou nutrição ao longo do tempo. Consequentemente, a informação não fica disponível para análise posterior (Silverstein, 2012).

Em qualquer empresa ou sistema de informação, considera-se que os dados têm qualidade quando estes cumprem os requisitos de precisão, completude, consistência, credibilidade, representação temporal e interpretabilidade (Han, Kamber, & Pei, 2012). Contudo, no mundo dos negócios continuam a existir muitos sistemas que não cumprem todos os requisitos, contendo quantidades significativas de dados incorretos, duplicados e que não representam a realidades (Caller, Pallat, & Darlow, 2008). Como conclui Olson (2003), estes dados originam atividades erradas, perda de clientes, perda de oportunidades e tomada de decisões incorretas, pelo que as organizações devem tomar consciência das consequências e tomar medidas para reverter esta tendência.

4. Data Mining

Vivemos num mundo onde são coletados, diariamente, quantidades massivas de dados. Este crescimento exponencial do volume de dados é resultado do rápido desenvolvimento tecnológico, acompanhado de ferramentas de recolha e armazenamento de dados. Para tirar o máximo valor deste recurso, são também necessárias ferramentas que permitam descobrir informações valiosas e transformá-las em conhecimento organizado e aplicável nas organizações (Han, Kamber, & Pei, 2012). Uma análise correta dos dados permite perceber o estado atual de um negócio, entender a evolução do que foi conquistado, mas também prever e antecipar ocorrências futuras. É por esta razão que empresas em todo o mundo registam todas as suas atividades de negócio, gerando fluxos de dados astronómicos (Han, Kamber, & Pei, 2012).

Na estratégia central de qualquer empresa de serviços está a habilidade de retenção e aquisição de novos clientes. Guo e Qin (2017), por exemplo, referem-se a melhorias significativas na qualidade do serviço prestado e na crescente orientação para a satisfação do cliente para as suas necessidades, como consequência da aplicação de técnicas de *data mining* em dados de CRM. O *data mining* surge assim como um elemento revolucionário na forma como o CRM é abordado nas organizações, desempenhando um papel determinante na tomada de decisão, no posicionamento competitivo das empresas e na concretização dessa visão estratégica (Guo & Qin, 2017). Neste contexto, o *data mining* apresenta-se como um conjunto iterativo de processos, que visa a compreensão, tratamento e processamento de grandes quantidades de dados. Esses processos são facilitados por ferramentas de *machine learning* que permitem a extração de informação relevante e significativa para o negócio (Hu, 2005). No presente capítulo apresentam-se os processos, técnicas e aplicações do *data mining*, assim como os princípios básicos para a construção de um modelo preditivo.

4.1. Introdução ao Data Mining

O termo *Data Mining* (mineração de dados) utiliza a expressão *mining* para caracterizar a descoberta de algo valioso na exploração de conjunto enorme de material em bruto, que neste contexto se materializa na descoberta de informação num conjunto de dados em bruto (Han, Kamber, & Pei, 2012). Han e Kamber (2001) definem *data mining* como uma etapa imprescindível no processo de Descoberta de Conhecimento em Base de Dados (do inglês *Knowledge Discovery in Databases* – KDD), no qual ocorre a descoberta de padrões relevantes e extração de informação útil a partir de uma quantidade enorme de dados provenientes de várias fontes, sejam estas relacionais, transacionais, orientadas a objetos,

espaciais, temporais, textuais, entre outras (Han & Kamber, 2001). Os padrões obtidos são utilizados para descrever conceitos, analisar associações, construir modelos de classificação e de regressão, identificar *clusters* de dados, modelar tendências em séries temporais ou detetar outliers, que representam dados que não cumprem o comportamento geral dos dados (Han, Kamber, & Pei, 2012). No conjunto dos vários processos, o objetivo final do *data mining* prende-se com a deteção de relações implícitas existentes entre os dados que, quando aplicadas ao contexto de negócio, auxiliam e acrescentam valor no momento da tomada de decisão (Zhao, Wu, & Gao, 2008; Shaoling & Yan, 2008).

Em concordância com estas definições, Hand, Mannila e Smyth (2001), Mehrotra e Agarwal (2009) e Guo e Qin (2017) definem *data mining* como uma área multidisciplinar que visa a extração de conhecimento, modelos e regras potencialmente valiosas de uma quantidade massiva de dados. Linoff & Berry (2011) acrescentam ainda que este processo não se resumo à descoberta de padrões nos dados, mas a padrões e regras realmente úteis e significativas para o negócio. Mais uma vez, os autores concordam que, ao calcular tendências e comportamentos futuros, estas técnicas promovem a tomada de decisão com base em conhecimento implícito nos dados. Enquanto um processo avançado de tratamento de dados, o *data mining* difere da análise de dados tradicional pois valoriza a exploração de relações e padrões entre os dados sem assumir premissas ou modelos estatísticos (Guo & Qin, 2017), integrando diferentes tecnologias e automatizando a extração do conhecimento (Greenyer, 2000).

Para as empresas, o *data mining* assume um papel diferenciador em diferentes áreas de negócio. Mehrotra e Agarwal (2009) referem que estas técnicas são maioritariamente utilizadas em organizações de *Business Intelligence*, serviços de análises financeiras, sistemas de informação e controlo de processos. Contudo, em literatura mais recente, são também muitos os exemplos de casos de sucesso em áreas de marketing. Guo e Qin (2017) aplicam técnicas de *data mining* para identificar os principais fatores que influenciam o comportamento de compra dos clientes, e assim, construir um modelo preditivo para as próximas interações. Ainda em 2008, Shaoling e Yan (2008), por exemplo, descrevem um novo modelo de gestão da relação com o cliente baseado em modelos preditivos.

Como afirmam Zhao, Wu e Gao (2008), o recurso ao *data mining* tornou-se quase obrigatório como consequência dos desenvolvimentos dos sistemas de informação que, ao produzirem e armazenarem quantidades massivas de dados, tornam insuficiente a capacidade analítica humana. Para reforçar a importância destes processos, Moody e Walsh (1999) enfatizam a dificuldade de produzir informação e de medir o seu impacto quantitativamente. Como referem os autores, um conjunto de dados em bruto não acrescenta nenhum valor por si só, sendo necessários vários processos de transformação e análise que consigam extrair informações úteis. A informação torna-se, então, um dos recursos mais valiosos e desejados nas organizações.

Como apresentado ao longo deste capítulo, o *data mining* foca tanto na integração e exploração dos dados, como na partilha de informação. A automação dos processos de transformação e modelação facilitam que a informação esteja constantemente disponível e pronta para consumo no momento correto. Em certas fases do processo, procede-se também à seleção criteriosa de dados e variáveis, pois como foi demonstrado, por vezes o excesso de informação também dificulta a sua análise e compreensão. Assim, pode concluir-se que a aplicação destes métodos permite uma maior orientação para o cliente e para a otimização de estratégias (Shaoling & Yan, 2008), assim como, simultaneamente, serve o propósito de aumentar o lucro das empresas (Mehrotra & Agarwal, 2009).

Na literatura existem diferentes abordagens para o encadeamento ótimo de processos que conduzem à descoberta de conhecimento nas bases de dados. No referente trabalho apresentam-se as duas metodologias mais aceites na literatura do *data mining*: *Cross Industry Standard Process for Data Mining* (CRISP-DM) e SEMMA (*Sample, Explore, Modify, Model, Assess*). Na fase de definição do problema, é essencial identificar o tipo de problema que se pretende abordar, pois consoante a questão a resolver, são aplicadas técnicas de pré-processamento e modelação distintas (Chapman, et al., 2000). De acordo com Chitra e Subashini (2013) existem duas grandes técnicas de *data mining*: Modelação Preditiva e Modelação Descritiva.

As técnicas de modelação preditiva, também conhecidas como *Supervised Learning* (aprendizagem supervisionada), baseiam-se na utilização de dados históricos, algoritmos estatísticos e de *machine learning* para identificar a probabilidade ou propensão de acontecimentos futuros (Ling & Li, 1998). Denomina-se por aprendizagem supervisionada pois é definida uma variável alvo (*target*), a qual é o objetivo da previsão. Estes algoritmos utilizam os dados cujos resultados são conhecidos para desenvolver e treinar o modelo, e assim, aprender os critérios de decisão que permitem classificar e prever o valor da variável *target* para novos registos (Linoff & Berry, 2011). Este tipo de técnicas pode ser dividido em dois grupos, consoante o tipo da variável *target*, sendo eles Classificação e Regressão.

Por outro lado, as técnicas de modelação descritiva são também conhecidas por *Unsupervised Learning* (aprendizagem não supervisionada), uma vez que não se aplica o conceito de variável alvo. Este tipo de técnicas explora as relações entre as várias variáveis e permite descrever e inferir padrões que não são facilmente detetáveis. Segundo Linoff e Berry (2011, pp. 81-88) são essencialmente utilizadas quando se pretende descrever, resumir, agrupar e compreender a organização de certas características, destacando-se as seguintes abordagens: Análise de Clusters, Regras de Associação e Visualização (Lin, Wan, & Pu, 2010). No contexto da presente dissertação, são apenas apresentadas com maior detalhe as técnicas de Classificação.

4.2. Aplicação de técnicas de Data Mining na indústria das telecomunicações

Com o aparecimento da internet, o crescimento exponencial de dados de clientes no sector móvel tornou-se o principal desafio para as empresas do sector. No estudo desenvolvido por Guifang e Youshi (2010), os autores abordam a necessidade de aplicação de técnicas de *data mining* no sector das telecomunicações, mais especificamente nas áreas de gestão de relação com o cliente, como resposta ao volume de dados produzido diariamente e à apertada concorrência. Linoff e Berry (2011) reforçam esta ideia, acrescentando que é fundamental reforçar a aprendizagem no que diz respeito às características, necessidades e comportamentos dos clientes, assim como interações e transações anteriores. Este conhecimento, aplicado em prol do negócio, pode ser um fator diferenciador tanto no aumento de receita, como na retenção ou aquisição de clientes. Apesar da tecnologia promover esta vantagem competitiva, na opinião de Chen e Ching (2007), trata-se apenas de uma vantagem temporária, pois à medida que a tecnologia evolui e vai sendo consolidada, torna-se também disponível para todos, deixando de surtir o seu efeito diferenciador. Segundo os mesmos autores, uma melhor abordagem passa por encarar a tecnologia como uma ferramenta potenciadora de uma estratégia focada no cliente, na qual o negócio deve focar-se em reter atuais clientes e aumentar a sua fidelização.

Na sua investigação sobre os fatores que impulsionam a fidelização dos clientes aos operadores de telecomunicações, Lin, Wan e Pu (2010) determinam que a lealdade para com uma marca pode expressar-se em dois nível: a nível comportamental ou atitudinal. A nível comportamental, a lealdade do cliente expressa-se através da sua experiência com os produtos e serviços, da combinação entre aqueles que são os seus interesses e o respetivo consumo de serviços. A nível atitudinal, a lealdade é avaliada de uma perspetiva psicológica e contempla a satisfação do cliente pela qualidade do serviço, pela garantia e confiança na marca e identificação pessoal com os valores e princípios transmitidos. Como resultado do seu estudo, os autores identificam que, para este sector, os principais fatores de fidelização de clientes são as relações de confiança e a satisfação do cliente a nível atitudinal, e as barreiras na migração entre operadores e as alterações de custos a nível comportamental (Lin, Wan, & Pu, 2010).

Na literatura, Lu, et. al. (2014) endereçam os temas de retenção e fidelização de clientes na industria das telecomunicações aplicando metodologias de *data mining.* Os autores desenvolveram um modelo de previsão de risco de portabilidade de clientes, isto é, um modelo que calcula a probabilidade de um cliente desativar o seu serviço e passar para uma operadora concorrente. O seu estudo indica que cerca de 2,2% dos clientes cancelam os seus serviços de telecomunicações por mês. Isto significa que 1 em cada 50 clientes muda de operadora todos os meses. Na perspetiva das operadores é extremamente mais rentável reter clientes existentes do que atrair novos, o que reforça a importância deste tipo de

mecanismos (Lu, Lin, Lu, & Zhang, 2014). Com este tipo de modelos, as empresas podem identificar potenciais clientes de risco, compreender os motivos que o levam a sair, e torná-los alvos de ofertas mais atrativas. Nestes casos, a colaboração com a área de CRM irá permitir interagir com o cliente no tempo certo, com a oferta correta, através do canal mais eficaz (Shaoling & Yan, 2008).

Outro exemplo de aplicação destas metodologias com efeitos na retenção, satisfação do cliente e na confiança e preferência pela marca foi apresentada por Pallegedara, et. al. (2006). Neste estudo, os autores desenvolvem um modelo que, com base nos padrões de consumo, prevê se os clientes vão ultrapassar o limite subscrito, gerando situações de falta de pagamento, que afetam a satisfação do cliente e que normalmente desencadeiam o processo de portabilidade para outros fornecedores. Prevenindo estas situações, as empresas são capazes de apresentar ofertas personalizadas e que melhor se adequam às necessidades dos clientes (Pallegedara, Amaratunga, Gopura, & Jayathileka, 2006). Esta é uma forma de antecipação dos problemas do cliente que tem um impacto muito grande na satisfação e na relação de confiança com o cliente.

Na opinião de Guifang e Youshi (2010), no mercado das telecomunicações impõe-se ainda a rentabilização das instalações e das infraestruturas de rede, que representam a maior fonte de custos fixos destas empresas. Este é outro fator que torna essencial a correta análise e previsão de mercado. Para os autores, mais do que garantir a retenção e aquisição de novos clientes, as empresas de telecomunicações devem também procurar desenvolver soluções flexíveis. Dada a facilidade e rapidez com que mudam as ofertas e exigências do mercado, as empresas devem trabalhar na sua capacidade de resposta, agilidade e ainda apostar em fatores de criatividade e personalização. Para se diferenciar de outros operadores e construir uma base de clientes leal, um fornecedor de serviços móveis deve atuar para além da tecnologia e apelar à individualidade dos seus clientes através de práticas de CRM (Chen & Ching, 2007).

Atualmente, dadas as mais diversas fontes de dados disponíveis, as operadoras podem não só tirar partido dos dados que armazenam dos seus clientes, mas também de potenciais clientes que demonstram interesse através de um dado canal (Coe, 2004b). Como mencionado anteriormente, as *lead* são conjuntos de dados facultados proactivamente por um indivíduo, e que representam potencial interesse num produto ou serviço (Willis & Flo, 2016). Neste âmbito as técnicas de data mining podem ser usadas para perceber as necessidades de potenciais clientes, determinar o tempo ótimo para ser contactado, calcular a propensão para uma nova aquisição ou até mesmo aumentar o valor de receita de um atual cliente.

Em suma, as tecnologias de *Data Mining* são inovadoras, no sentido em que apresentam diversas ferramentas analíticas que se complementam e permitem explorar os dados armazenados, modelá-los e implicitamente, estabelecer regras e estratégias de negócio. A aplicação de *data mining* às práticas de CRM e ao suporte da tomada de decisão vem alavancar a aquisição de novos clientes, o aumento de valor dos clientes atuais, a retenção dos clientes mais valiosos e a oferta de serviços cada vez mais customizados, o que, consequentemente, coloca a empresa numa posição competitiva mais favorável. (Guifang & Youshi, 2010). Como referem Zhao, Wu e Gao (2008), ao aplicar estas metodologias ao CRM, as empresas de telecomunicações estão, simultaneamente, a lidar com todas a informação que conseguem recolher sobre os clientes, assim como a extrair novo conhecimento sobre os seus consumidores, retirando benefícios diretos para qualidade do seu serviço e para a estratégia competitiva da empresa.

4.3. Modelação Preditiva

As técnicas de modelação preditiva, também conhecidas como *Supervised Learning* (aprendizagem supervisionada), baseiam-se na utilização de dados históricos, algoritmos estatísticos e de *machine learning* para identificar a probabilidade ou propensão de acontecimentos futuros. A supervisão da aprendizagem é obtida através dos registos previamente classificados do conjunto de dados de treino (*training dataset*). À variável que determina a que classe pertence cada linha do *training set* chama-se variável alvo ou *target* (Han, Kamber, & Pei, 2012). Estes algoritmos utilizam os dados cujos resultados são conhecidos para desenvolver e treinar o modelo, e assim, "aprender" os critérios de decisão que permitem classificar e prever o valor da variável *target* para novos registos. Este tipo de modelação difere da modelação descritiva, muitas vezes denominada por Análise de Clusters, na qual não é atribuído nenhuma classe a cada um dos registos na fase de treino, e no qual o número de classes resultantes da aprendizagem não é conhecido previamente.

Como mencionado anteriormente, este tipo de técnicas pode ser dividido em dois grupos, consoante o tipo de variável alvo. Em problemas cuja variável alvo é binária, isto é, só admite dois valores, são aplicadas técnicas de Classificação. No caso em que o campo a ser previsto admite valores contínuos, aplicam-se técnicas de Estimativa ou Regressão. Na presente dissertação foram abordados com maior detalhe as técnicas de Classificação.

No processo de modelação preditiva, incorporado no processo de Descoberta de Conhecimento em Base de Dados, podem ser adotadas diferentes abordagens das quais CRISP-DM e SEMMA, descritas nos capítulos seguintes.

4.3.1. *Cross Industry Standard Process for Data Mining* (CRISP-DM)

A sigla CRISP-DM do inglês *Cross Industry Standard Process for Data Mining*, é uma metodologia aplicada em várias indústrias para aumentar o sucesso de projetos de *data mining.* Esta metodologia define uma sequência flexível de seis processos, a qual permite a construção e implementação de um modelo de *data mining* à medida de problemas reais, com o objetivo de responder às necessidades das empresas e suportar, diariamente, as decisões de negócio (Chapman, et al., 2000). Segundo Witten e Frank (2005) a grande vantagem deste processo é o seu cariz iterativo que permite que o resultado final seja completamente afinado e alinhado com os objetivos de negócio.

Na figura 4.1 é apresentado o fluxo de processos do modelo CRISP-DM, como descrito por Chapman, et al. (2000). Na imagem, as setas indicam as dependências entre as fases que ocorrem mais frequentemente. O ciclo exterior representado na imagem simboliza o ciclo natural do próprio *data mining.* Uma vez que o processo de *data mining* não termina quando a solução é lançada, as lições aprendidas durante o processo e mesmo depois do lançamento devem ser novamente incorporadas no modelo. Assim, os processos subsequentes beneficiam das experiências e resultados anteriores.

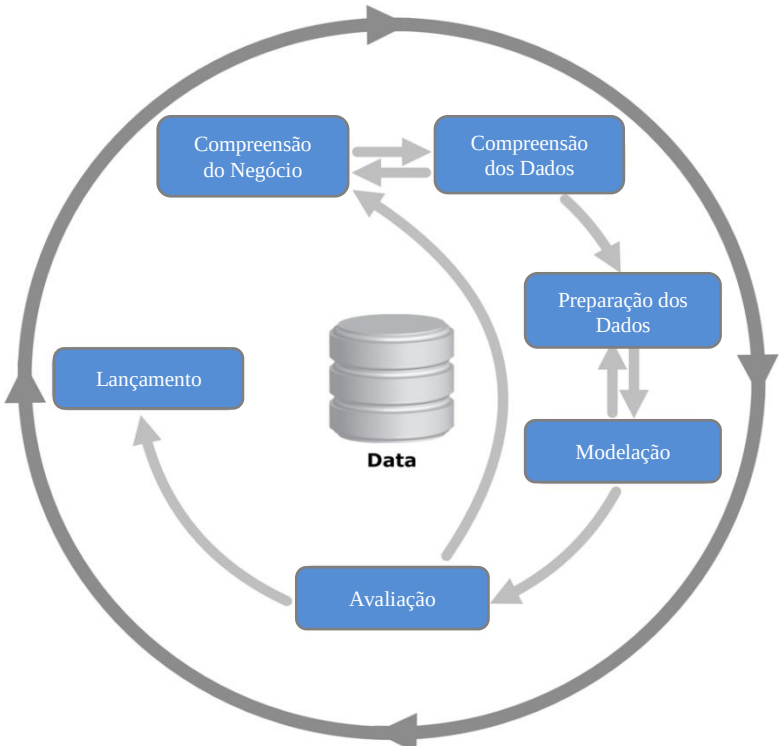

Figura 4.1 – Fases do modelo CRISP-DM (adaptado de Chapman, et al., 2000).

Segundo Chapman, et al. (2000), as várias fases do processo de *data mining* são as seguintes:

- **Compreensão do Negócio** – a fase inicial foca na compreensão dos objetivos do projeto e requisitos do ponto de vista de negócio. A fase seguinte passa por converter esse conhecimento na definição de um problema de *data mining* e num planeamento preliminar para alcançar esses objetivos.

- **Compreensão dos Dados** – esta fase começa com a recolha de dados e continua com atividades que permitem familiarizar-se com os dados, identificar problemas na qualidade dos mesmos, extrair as primeiras impressões sobre o conjunto de dados e detetar padrões interessantes.

- **Preparação dos Dados** – esta fase cobre todas as atividades necessárias para construir e obter o *dataset* final, o qual irá alimentar as ferramentas de modelação. Algumas das tarefas incluídas nesta fase são limpeza de dados, seleção de atributos, assim como transformação dos dados para a ferramentas de modelação.

- **Modelação** – nesta fase são selecionadas e aplicadas várias técnicas de modelação, e os respetivos parâmetros são afinados para valores ótimos. Como cada modelo tem requisitos específicos para o tipo de data que consegue processar, é por vezes necessário retroceder à fase de preparação e transformação dos dados.

- **Avaliação** – neste ponto do projeto, foram contruídos modelos que aparentam ter elevada qualidade numa perspetiva analítica. A avaliação dos modelos é efetuada comparando o seu desempenho através de várias métricas. No entanto, antes de proceder para o lançamento final da solução é essencial avaliar e rever todos o processo e certificar-se que o modelo responde aos objetivos estabelecidos inicialmente. Caso algum objetivo não esteja inteiramente concluído, o modelo permite sempre voltar ao início do processo. No final desta fase deverá ser tomada a decisão se os resultados do modelo estão prontos para ser aplicados.

- **Lançamento** – esta fase preconiza a organização e apresentação do conhecimento ganho, de uma forma que possa ser utilizada pelo cliente final. Dependendo dos requisitos, a fase de lançamento da solução pode ser tão simples como gerar um relatório, ou ser mais complexa como implementar um processo automáticos de *data mining* na empresa.

4.3.2. SEMMA

Adicionalmente ao método de CRISP-DM, outra metodologia referenciada na literatura, desenvolvida pelo SAS *Institute*, é o processo SEMMA (Olson & Delen, 2008), demonstrado na figura 4.2. Cada fase do processo é descrita a seguir:

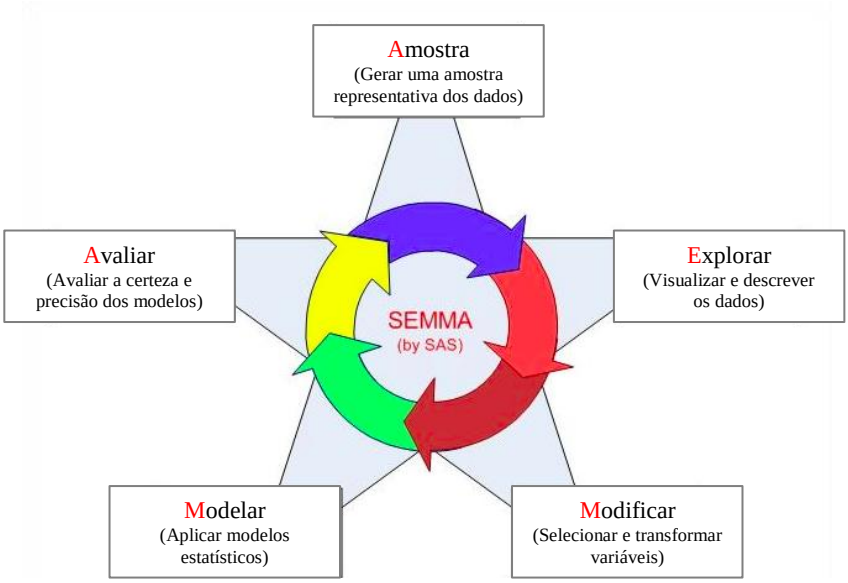

Figura 4.2 - Processo SEMMA, *Sample*, *Explore*, *Modify*, *Model*, *Assess* (adaptado de Olson & Delen, 2008).

- **Amostra** – nesta fase são extraídas amostras representativas dos dados, de modo a melhorar o desempenho computacional e a reduzir o tempo de processamento. É também nesta fase que ocorre a partição dos dados em conjuntos de treino, validação e teste.

- **Explorar** – através da fase exploração e visualização são antecipadas alguma relações, tendências e anomalias nos dados, facilitando a compreensão do objeto em estudo e assegurando a qualidade dos dados.

- **Modificar** – como resultado das descobertas feitas na fase de exploração, pode ser necessário excluir, criar ou transformar alguns registos ou variáveis antes da fase de modelação. É também importante verificar a presença de outliers, os quais, se não forem eliminados, podem prejudicar o desempenho e o poder preditivo do modelo.

- **Modelar** – nesta fase do projeto é desempenhada a procura do modelo que melhor prevê a variável pretendida, respondendo ao objetivo estabelecido. Consoante o problema, podem ser aplicados dois grupos de modelos. O grupo dos modelos descritivos, também conhecido como

Unsupervised Learning, é um conjunto de técnicas que descreve a estrutura e resume os dados. O grupo dos modelos preditivos, também conhecido como *Supervised Learning*, tem o objetivo de criar estruturas que permitam prever com algum grau de confiança o resultado de um evento, baseado num conjunto de dados previamente classificados.

- ***Assess* (Avaliar)** – Nesta última fase do processo são avaliadas as várias alternativas de modelos e é comparado o desempenho de cada um de acordo com várias métricas. Tipicamente, cada modelo é aplicado a um conjunto de dados que não foi utilizado na sua construção, sendo assim possível analisar uma estimativa não enviesada do seu poder preditivo.

Os dois processos de *data mining* mencionados dão uma visão geral das várias fases de desenvolvimento de um modelo preditivo. Foram abordados ambos os processos pois, embora existam algumas semelhanças, a metodologia CRISP-DM descreve o processo num contexto de negócio, enquanto que o SEMMA detalha os procedimentos técnicos necessários para construir o modelo, uma vez definidos os objetivos de negócio. Uma vez apresentados os conceitos básicos referentes às técnicas de *data mining*, serão de seguida descritos com maior detalhe os procedimentos da construção de um modelo preditivo, aplicados na secção prática.

4.3.3. Classificação

A classificação é um dos métodos mais usados do *data mining*, uma vez que é quase inato da condição humana. Este consiste em atribuir classes ou categorias a um novo acontecimento, com base em experiências anteriores. O processo de classificação é caracterizado por um conjunto de dados previamente classificados, um conjunto de classes pré-definido e um modelo capaz de aprender corretamente os critérios de classificação (Han, Kamber, & Pei, 2012). O objetivo é construir um modelo que, quando aplicado a dados novos e não classificados, seja capaz de atribuir uma classificação acertada (Linoff & Berry, 2011). Neste método, o modelo é construído e derivado dos dados históricos, sendo também designado por Classificador, podendo ser apresentado de diversas formas, como regras de classificação, árvores de decisão, redes neuronais, entre outros (Han, Kamber, & Pei, 2012).

Como foi apresentado na secção 4.3, os métodos de classificação distinguem-se dos de regressão por preverem apenas categorias de valores, isto é, variáveis categóricas, discretas e não ordinais. Os métodos de regressão, por sua vez, permitem uma previsão numérica de uma variável, admitindo valores contínuos ou ordinais. Na sua obra, Han, Kamber e Pei (2012) identificam duas fases distintas no processo de classificação: a **fase de aprendizagem** e a **fase de classificação**. Ao desenvolver um modelo

preditivo, é fundamental assegurar que o modelo é validado e testado com dados complementares aos que foram utilizados durante a fase de aprendizagem. Para tal, são produzidos três conjuntos de dados:

- ***Training Data*** (dados de treino) – conjunto de dados com um valor pré-definido para a variável *target*, utilizados para construir o modelo preditivo, ou classificador. Quanto maior o *training dataset*, melhor a qualidade do modelo obtido.

- ***Validation Data*** (dados de validação) – conjunto de dados, também previamente classificados, utilizados para uma avaliação inicial do poder preditivo do modelo, afinando os algoritmos e evitando a aprendizagem excessiva dos dados, limitando a fase de aprendizagem. Quanto maior o *validation dataset*, maior a confiança na fase de treino.

- ***Testing Data*** (dados de teste) – conjunto de dados previamente classificados, completamente independentes da construção e afinação do modelo. Este conjunto de dados é utilizado para a avaliação final do modelo, fornecendo uma estimativa real e um nível de confiança certo para o poder preditivo do modelo em dados desconhecidos. Quanto maior o *testing dataset*, melhor a estimativa de *performance* do modelo em dados reais.

Na figura 4.3 é apresentado um esquema da estrutura básica de um modelo preditivo e as respetivas funções de cada conjunto de dados.

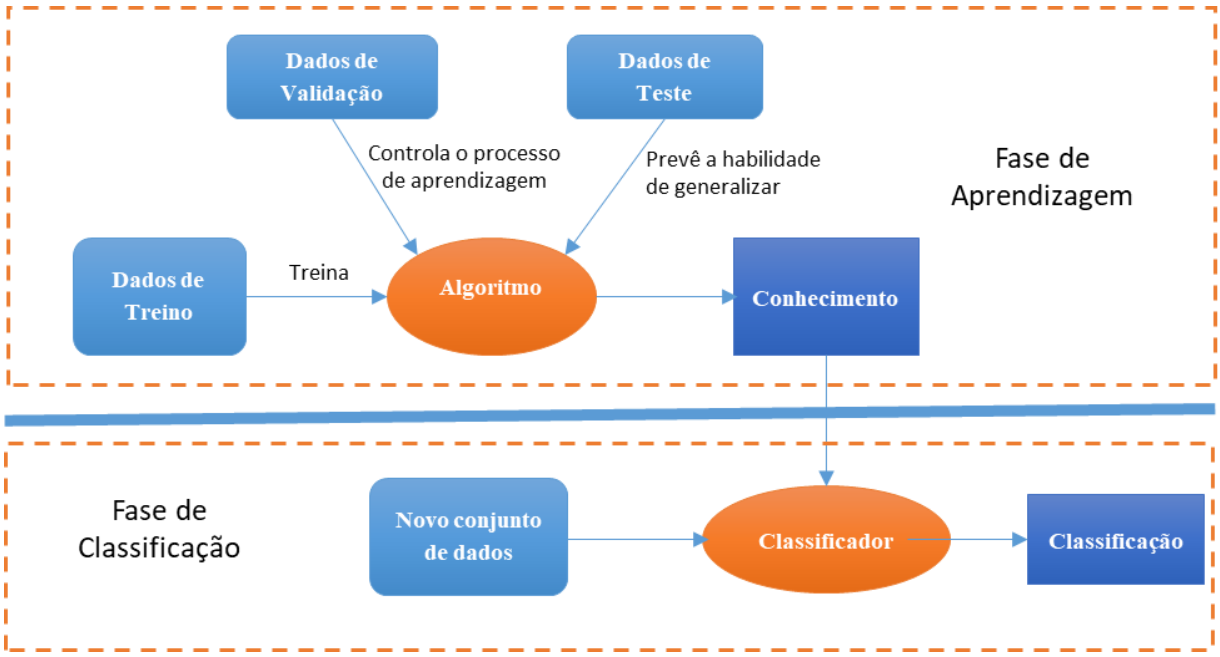

Figura 4.3 - Estrutura de um modelo preditivo (adaptado de Han & Kamber, 2001)

Na fase de aprendizagem, os dados de treino são utilizados para treinar e ensinar o modelo com base em acontecimentos e padrões anteriores, enquanto que os dados de validação são utilizados simultaneamente para efetuar ajustes na parametrização do algoritmo. Na fase de classificação, o modelo é utilizado para prever e classificar novos registos. Como indicado na figura 4.3, os dados de teste permitem estimar a precisão preditiva do classificador quando aplicados a dados desconhecidos. Se a precisão fosse estimada com base nos dados de treino, o resultado obtido seria extremamente otimista, uma vez que o classificador tem tendência para aprender em demasia os dados, isto é, pode memorizar alguns fenómenos ou anomalias que não seguem o padrão genérico do conjunto de dados. A este acontecimento designa-se por *overfitting*, que significa precisamente sobreaprendizagem dos dados. Assim, os dados de teste garantem que o modelo é testado e avaliado em dados independentes e não enviesados.

Como sugerem Hand, Mannila e Smyth (2001), no desenvolvimento de um modelo preditivo, o conjunto de dados inicial é dividido nos 3 conjuntos, segundo as proporções 60% - 20% - 20% para os dados de treino, validação e teste, respetivamente. No subcapítulo seguinte é feita uma breve apresentação do problema de *overfitting* na construção de modelos preditivos e potenciais causas e efeitos.

4.3.3.1. O Problema de *Overfitting*

Conforme abordado anteriormente, as técnicas de *data mining* têm como objetivo ajustar um modelo a um conjunto de dados de treino, com o intuito de fazer previsões fiáveis em dados desconhecidos. Este objetivo requer que o algoritmo analise e aprenda padrões de comportamento e relações entre as variáveis.

A sobreaprendizagem consiste no ajustamento excessivo do modelo ao conjunto de treino, dificultando a realização de boas classificações perante novos dados (Hand, Mannila, & Smyth, 2001). Quando ocorre sobreaprendizagem significa que o modelo descreve os dados de treino ao pormenor, memorizando erros aleatórios e ruídos nos dados, em vez de descrever as relações subjacentes. Este problema ocorre quando um modelo é excessivamente complexo, por exemplo, quando tem demasiados parâmetros face ao número de observações consideradas. Nesta situação, o modelo começa a "memorizar" o conjunto de dados de treino em vez de desenvolver a capacidade de generalizar tendências. Deste modo, um modelo com estas características tem um fraco poder preditivo, uma vez que reage a flutuações menores nos dados de treino. Na figura 4.4 apresenta-se um caso simples de sobreaprendizagem do modelo aos dados.

Considere-se, por exemplo, o conjunto de dados de treino, representados por círculos, os quais podem ser classificados em duas categorias, azuis ou vermelhos.

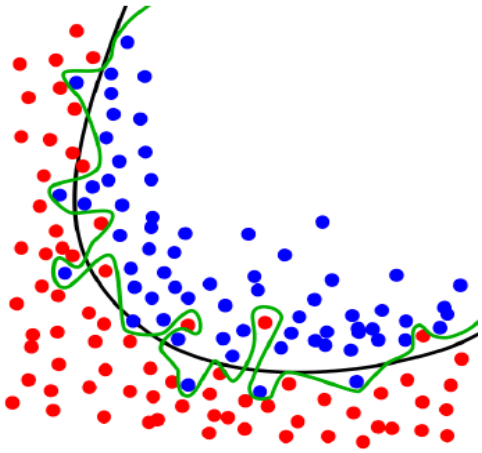

Figura 4.4 - Demonstração do problema de sobreaprendizagem na modelação preditiva.

Um exemplo de um bom modelo preditivo seria a função desenhada a preto, a qual classificaria um registo na parte interior da curvatura como azul, e na parte exterior como vermelho. Um exemplo de um modelo *overfitted* seria o traçado a verde, o qual claramente memorizou todo o conjunto de dados, respondendo às várias flutuações de valores e perdendo a capacidade de generalização do modelo. O problema pode ocorrer porque os critérios de treino de um modelo não são os mesmos critérios utilizados para testar a precisão do modelo (Han & Kamber, 2001). Tipicamente, na fase de treino do modelo, o objetivo é maximizar o número de registos classificados corretamente. No entanto, a precisão final do modelo é determinada com base na sua habilidade de classificar corretamente em instâncias que não foram previamente consideradas. Na figura 4.5 é representado o ponto ótimo de aprendizagem de um modelo, de modo a evitar situações de *overfitting*.

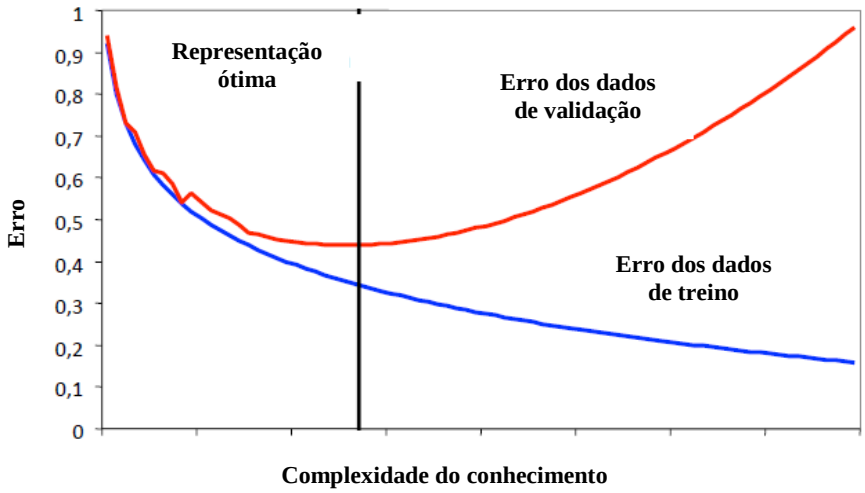

Figura 4.5 - Representação das curvas de erro nos conjuntos de treino e validação (Adaptado de Kurita, 2018) .

Como se pode observar na figura 4.5, o erro associado ao processo de aprendizagem é inversamente proporcional à complexidade do próprio processo, decrescendo a uma taxa menor à medida que a complexidade aumenta. O erro associado à validação, por sua vez, sofre um decréscimo no início do treino, contudo, inverte esta tendência a certo ponto, aumento quando exponencialmente com a complexidade da representação (Kurita, 2018). Isto significa que, a certa altura do processo de aprendizagem, quanto maior a complexidade do modelo, menor a precisão na classificação de eventos novos. Esta situação designa-se por *overfitting*, pois os dados descrevem absolutamente os dados de treino. Assim, considera-se o ponto ótimo de aprendizagem aquele onde ocorre a inversão do declive da curva de erro do conjunto de validação (Kurita, 2018).

4.4. Processamento de Dados

Atualmente, as bases de dados são altamente suscetíveis a inconsistências, falhas e ruídos nos dados. Este facto deve-se maioritariamente à diversidade heterogénea de fontes de dados, assim como, à dimensão e volume de dados armazenados (Han, Kamber, & Pei, 2012). Em muitas empresas, existe também a particularidade de muitos dados serem atualizados ou preenchidos manualmente, adicionando variabilidade aos valores recolhidos e potenciando a ocorrência de erro (Zhao, Wu, & Gao, 2008).

Han, Kamber e Pei (2012) refletem sobre a importância do pré-processamento dos dados no processo de *data mining*, uma vez que dados com pouca qualidade produzem resultados com pouca qualidade. Os autores identificam 6 fatores que compõem a definição de qualidade de dados, sendo eles a precisão, completude, consistência, credibilidade, representação temporal e interpretabilidade. Os autores evidenciam que, independentemente da fonte de dados, os dados que alimentam o processo de *mining* devem ser previamente processados de modo a reduzir o efeito de ruídos, valores em falta, inconsistências e erros na fase de modelação. Embora seja a fase mais morosa do projeto, a correta limpeza, integração, tratamento e transformação dos dados é essencial para obter um modelo preciso (Hu, 2005).

No seguimento deste capítulo são apresentadas várias técnicas de processamento de dados que permitem melhorar a qualidade dos mesmos, realçar padrões existentes e reduzir o tempo de processamento do modelo. Os métodos de processamento de dados estão organizados nas seguintes categorias: limpeza dos dados, transformação, partição e redução de dimensionalidade.

4.4.1. Limpeza dos Dados

No contexto empresarial e industrial, a limpeza dos dados é um problema recorrente. No entanto, só tem sido alvo de investigação e publicações nas últimas décadas. Rahm e Hai Do (2000) identificam como principais razões o facto de ser um tarefa extremamente complexa e intensiva. Como identificam os autores, bases de dados heterogéneas, com diferentes estruturas e semânticas, múltiplos canais e plataformas de captura de dados dentro da mesma empresa, são alguns dos desafios desta área. Apesar dos esforços para uniformizar o fluxo de dados, continua a ser essencial a limpeza e organização prévia, de forma reduzir a complexidade e os custos de processamento (Rahm & Hai Do, 2000).

Adicionando a "variável *internet"*, torna-se ainda maior a proliferação de dados a considerar, e mais relevante e desafiante a aplicação destas metodologias. No caso concreto da geração online de *Lead*, as bases de dados tendem a armazenar grandes volumes de dados heterogéneos e com muitos erros. Neste exemplo, para obter dados mais homogéneos e uniformes é essencial uma análise cuidada dos atributos e dos seus valores, seguida de métodos de limpeza (Gupta, Wasid, & Ali, 2016). Algumas das atividades chave passam pela identificação de registos duplicadas dentro do mesmo canal ou entre canais comerciais, ou pela a uniformização da estrutura dos dados armazenados. Por si só, estas medidas já permitem uma segmentação mais facilitada e uma modelação mais eficiente.

Han, Kamber e Pei (2012) definem 3 principais problemas dos dados, sendo eles *missing values* ou valores em falta, *outliers* (termo inglês para valores extremo com comportamento diferente dos restantes) e erros. Para cada um dos problemas pode ser adotado diversas abordagens de resolução, apresentadas de seguida.

Missing Values - correspondem a valores em falta, podendo ser expressados com um campo em branco ou preenchido com um valor por defeito, por exemplo, "*null"*. As diferentes formas de lidar com *missing values* são:

1. Remoção da variável – neste cenário há perda de informação. Só é aconselhável quando uma grande percentagem de registos apresenta valores por preencher para determinada variável;
2. Remoção do registo – neste cenário impõem-se padrões aos dados e perde-se informação das restantes variáveis;
3. Preenchimento manual com um valor provável – esta tarefa consome muito tempo e pode ser impossível face ao tamanho do conjunto de dados;
4. Preenchimento automático com uma medida de tendência central (média, mediana ou moda) – esta opção tende a diminuir o erro introduzido nos dados, sendo que a média deve ser utilizada

preferencialmente em variáveis com distribuição normal ou simétrica, enquanto que a mediana ou a moda devem ser aplicadas a distribuições assimétricas, uma vez que são menos sensíveis a *outliers;*

5. Preenchimento automático com uma métrica de tendência central de um subconjunto de dados – este exemplo segue a mesma lógica do anterior, com a particularidade de que a métrica aplicada é calculada num subconjunto de dados que tem características em comum com a variável que se pretende preencher;
6. Preenchimento com o valor mais provável – o valor mais provável pode ser determinado com uma regressão linear ou uma árvore de decisão;

Para cada caso em particular deve ser avaliado qual o método que melhor se adequa, e que adiciona menor erro ao processo. Em comparação com os restantes métodos, a opção 6 é aquela que tira partido da informação disponível nos dados para prever o valor em falta. Ao considerar outras variáveis na sua estimativa, há uma maior probabilidade de serem preservadas as relações entre as variáveis.

Outliers **-** Estatisticamente, Han, Kamber e Pei (2012, pp. 543-548) definem um *outlier* como uma observação que está distante de outras observações, e que, por ser um caso extremo, tem um grande impacto na interpretação dos resultados. Os *outliers* podem ser produzidos em medições incorretas, erros na recolha de dados ou valores corretos que realmente fogem do padrão normal dos dados. Os autores descrevem várias abordagens para a deteção de *outliers*, como por exemplo métodos estatísticos, baseados na proximidade, baseados na densidade, análise de *clusters*, métodos paramétricos, não-paramétricos e outras variantes (Han, Kamber, & Pei, 2012, pp. 549-573). Contudo, em primeira instância a deteção pode ser efetuada recorrendo a métodos gráficos que, para além de mais simples e intuitivos, produzem resultados mais rapidamente. Das várias técnicas gráficas para identificar *outliers* destacam-se as seguintes:

1. **Histograma** – permite observar a dispersão de indivíduos pela gama de valores;
2. **Gráfico de dispersão bidimensional** – permite determinar *outliers* entre pares de variáveis;
3. **Box-Plot** – o gráfico distribui os dados em quartis, identificando os limites máximo e mínimo para o intervalo de dados. Os valores que ultrapassem esses limites são considerados *outliers.*

Após a identificação de valores extremos, esses valores devem ser avaliados e discutidos para perceber se se trata realmente de um erro ou de um valor legítimo, mas diferente dos restantes. Tendo em consideração que os métodos estatísticos são sensíveis a perturbações nos dados e que a presença de *outliers* pode produzir resultados instáveis, deverá ser tomada uma decisão que poderá passar por

remover o registo do conjunto de dados, ou até ser imposto um limite superior e inferior para cada variável.

Erros - A deteção de erros ou anomalias ocorre na mesma análise da deteção de *outliers*, chegando à conclusão que o valor foi incorretamente inserido na base de dados, seja por medições incorretas, introdução manual, vírgulas ou zeros mal colocados. Nestes casos, pode optar-se por corrigir localmente ou remover o registo, uma vez que o total de registos filtrados não ultrapasse 3% do conjunto total (Han, Kamber, & Pei, 2012).

4.4.2. Transformação de dados

Na construção de um modelo preditivo, a qualidade do resultado é altamente condicionada pelas variáveis utilizadas, sendo da maior importância a correta seleção destas. As variáveis devem ser representativas da realidade, mas devem também estar alinhadas com o problema em estudo e com as necessidades de negócio. Na fase de transformação de dados, os dados são transformados ou consolidados de forma apropriada, assegurando os requisitos da fase de modelação. Segundo Han, Kamber e Pei (2012) as estratégias para transformação de dados incluem as seguintes:

1. **Minimização de Ruído** – remoção de ruído dos dados através de técnicas como regressão, *clustering* ou *binning* (termo inglês para a discretização dos dados, dividindo variáveis contínuas em intervalos).
2. **Construção de Variáveis –** construção de novas variáveis através de variáveis existentes, e adição das mesmas ao conjunto de dados.
3. **Agregação –** tipicamente é utilizado na construção de cubos de dados para análises a diferentes níveis de granularidade e detalhe.
4. **Normalização –** conversão dos dados para ajustarem numa escala única, geralmente com uma amplitude menor, tal como [-1;1] ou [0;1].
5. **Discretização –**conversão de variáveis numéricas em variáveis intervalares ou categóricas. A variável idade, por exemplo, pode ser transformada em intervalos como 0-10, 11-20, ect. ou dividia em classes como Jovem, Adulto e Senior. A discretização pode ser aplicada recursivamente a vários níveis para criar uma hierarquia entre as variáveis.

As técnicas de transformação podem ser aplicadas com a função de realçar padrões nos dados, facilitando o processo de aprendizagem e modelação, mas é também aplicado na correção de inconsistências.

4.4.3. Redução de Dimensionalidade – Métodos de seleção de Variáveis

A identificação de dados relevantes, na quantidade certa e referente a um período temporal significativo é crítico para a qualidade do modelo desenvolvido (Hu, 2005). A redução de dimensionalidade é o processo de redução do número de variáveis ou atributos considerados, utilizado na fase de processamento dos dados em *data mining*. O objetivo prende-se com a melhoria do desempenho do algoritmo utilizado, potenciando a precisão da previsão e a interpretação dos resultados (Wang, Sanguansintukul, & Lursinsap, 2008; Kumar, Kongara, & Ramachandra, 2013).

Na análise de dados pode ocorrer um fenómeno designado por "*The curse of dimensionality"*, também conhecido como "a maldição da dimensionalidade". Na realidade, a existência de muitas variáveis pode ter ambos os efeitos, ser uma "bênção" ou uma "maldição". Geralmente, mais variáveis significam mais informação descritiva disponível, informação essa que pode ser aproveitada e utilizada para construir modelos melhores. Por outro lado, como referem Linoff e Berry (2011), muitas variáveis podem representar um problema por várias razões: elevada correlação entre as variáveis; aumento do risco de *overfitting* e maior dispersão dos dados.

Tipicamente, quanto maior o número de variáveis, maior a probabilidade de haver um elevado grau de correlação entre elas. Como referem Linoff e Berry (2011), essas relações entre as variáveis afetam negativamente os resultados da modelação. Algoritmos como Regressões, Redes Neuronais e Árvores de Decisão não funcionam tão bem com muitas variáveis, pois quando existem variáveis que descrevem o mesmo fenómeno, dificultam a extração de padrões dos dados (Linoff & Berry, 2011). Para além disso, variáveis altamente correlacionadas acrescentam a mesma informação ao modelo, tornando redundante a sua utilização, e aumentam o tempo de processamento do modelo desnecessariamente.

O risco de sobreaprendizagem está presente em muitas técnicas de *data mining*. Contudo, um número elevado de variáveis potencia as oportunidades para a ocorrência deste fenómeno em técnicas como Regressão e Redes Neuronais. Nestes algoritmos, a adição de variáveis ao modelo resulta na adição de graus de liberdade ao próprio modelo, o que facilita a memorização dos dados em vez da generalização. Por esta razão, é importante reduzir o número de variáveis quando aplicadas estas técnicas.

Por último, quando a dimensionalidade aumenta, o espaço de valores torna-se mais disperso, tornando-se mais difícil encontrar grupos e padrões relevantes entre os indivíduos. Observe-se a figura 4.6, onde se demonstra com um exemplo simples o efeito da dimensionalidade.

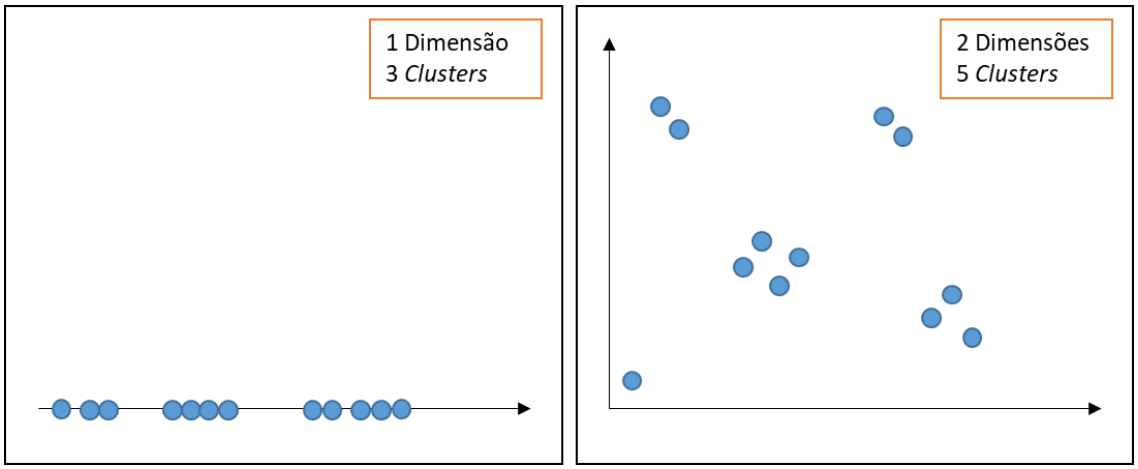

Figura 4.6 - Projecção Unidimensional e Bidimensional dos dados.

Analisando a figura à esquerda, o conjunto de valores distribui-se apenas numa variável e é possível identificar três grupos distintos. Adicionando uma variável à representação, a dispersão dos dados no espaço é maior, o que torna mais difícil a deteção de *clusters* e de semelhanças entre os indivíduos. Escalando este exemplo para um *dataset* real, com milhares de registos e centenas de atributos, torna-se extremamente complexa a modelação e interpretação dos dados. Assim, pode afirmar-se que a redução de dimensionalidade garante o valor e a validade do modelo preditivo (Guo & Qin, 2017).

As técnicas de redução de dimensionalidade podem ser aplicadas para obter uma representação reduzida do conjunto de dados, contudo, mantendo a integridade do conjunto original. Variáveis irrelevantes e redundantes são identificadas e removidas do conjunto de dados. O risco de manter estas variáveis em análise pode resultar na confusão do algoritmo aplicado e na descoberta de padrões de baixa qualidade. Estas técnicas têm um impacto positivo na eficiência do modelo, produzindo praticamente os mesmos resultados estatísticos. Nesta secção são apresentadas algumas das estratégias de seleção de atributos que incluem *Principal Components Analysis (*Análise de Componentes Principais) e Métodos de Regressão, praticadas por Han, Kamber e Pei (2012).

4.4.3.1. Análise de Componentes Principais

A análise de componentes principais, conhecido por *Principal Components Analysis (PCA)* na literatura original, é um dos métodos alternativos para reduzir o número variáveis. Este método consiste na transformação matemática dos dados de *input* em vetores ortogonais, segundo combinações lineares entre as variáveis, procurando os k vetores n-dimensionais que melhor representam os dados, tal que $k \leq n$. Os dados originais são, assim, projetados num espaço muito menor, resultando na redução da

dimensionalidade (Han, Kamber, & Pei, 2012). Ao contrário de outros métodos que reduzem a dimensionalidade ao reterem subconjuntos de atributos do conjunto inicial, a análise de componentes principais combina matematicamente variáveis, criando um conjunto alternativo e reduzido de variáveis. No entanto, estas novas variáveis, designadas por Componentes Principais (*PC*), tornam-se de difícil interpretação direta por parte de um analista.

Os procedimentos básicos para aplicação da análise de componentes principais passam pela normalização dos dados de *input*, para que cada atributo seja distribuído dentro do mesmo intervalo. Este passo ajuda a garantir que atributos com grandes espectros de valores não dominam atributos com menores amplitudes. De seguida, são calculados k vetores ortogonais, também denominados por componentes principais, que resultam da combinação linear dos dados normalizados. Essencialmente, os *PC* representam novos conjuntos de eixos perpendiculares, fornecendo informação importante sobre a sua variância. Uma vez calculados, os *PC* são ordenados por ordem decrescente de significância, tal que o primeiro eixo representa a maior variância dos dados, o segundo eixo representa a segunda maior variância dos dados, e assim sucessivamente. Segundo Linoff e Berry (2011) o 1º componente principal é aquele que, de todas as combinações lineares possíveis entre as variáveis de *input*, maximiza a variância da projeção desses inputs numa linha. A representação geométrica, no entanto, é muito mais intuitiva para interpretar este conceito. A figura 4.7 exibe os dois primeiros *PC*, Y_1 e Y_2, para um dado conjunto de dados originalmente representados pelos eixos X_1 e X_2.

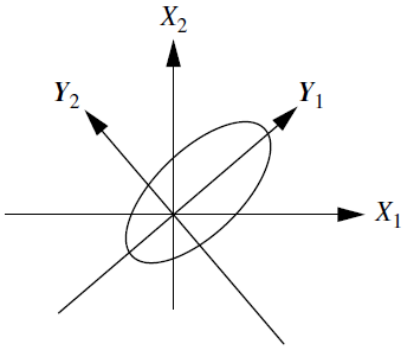

Figura 4.7 - Análise de Componentes Principais.

Como é possível observar na figura acima, o vetor Y_1 é representado na direção de maior variabilidade dos dados, enquanto que Y_2 na direção perpendicular. Esta informação ajuda a identificar grupos e padrões escondidos nos dados. Uma vez que os componentes são ordenados por ordem decrescente de significância, o tamanho do conjunto de dados pode ser reduzido ao eliminar os componentes mais fracos, isto é, aqueles com menor significância e a acrescentam menos variabilidade aos dados de *input*.

"Utilizando apenas os *PC* com maior significância é possível reconstruir uma boa aproximação do *dataset* original" (Han, Kamber, & Pei, 2012). Uma das particularidades deste método é que o número de *PC* é igual ao número inicial de variáveis e pode ser aplicado em qualquer conjunto de dados multidimensional, resolvendo problemas como dispersão e enviesamento dos dados.

4.4.3.2. Regressão Linear Múltipla

Os métodos de regressão integram as técnicas de seleção de subconjuntos de atributos ao reduzir o tamanho do conjunto de dados, removendo atributos irrelevantes e redundantes. O objetivo desta seleção é encontrar o conjunto mínimo de atributos tal que, a distribuição de probabilidades dos dados seja tão próxima quanto possível da distribuição original, obtida com todos os atributos (Blattberg, Kim, Kim, & Neslin, 2008).

No entanto, uma procura exaustiva pele subconjunto ótimo pode ser extremamente caro, principalmente à medida que o número de variáveis e de registos aumenta. Assim, a aplicação de métodos heurísticos, cujo objetivo é procurar no espaço de atributos pela escolha ótima local, com a espectativa que esta conduza à solução ótima global, pode simplificar esta tarefa. Na prática, estes métodos são eficazes e aproximam-se muitas vezes da solução ótima. Tipicamente, a escolha de atributos é definida recorrendo a testes estatísticos, onde é determinada a significância estatística de cada variável, e assumindo que os atributos são independentes entre si. Os principais métodos heurísticos para a seleção de atributos incluem técnicas de regressão como *Forward Selection* (Seleção para a frente), *Backward Elimination* (Eliminação para trás) e *Spetwise Selection* (Seleção Passo-a-Passo), que resulta da combinação das duas anteriores (Linoff & Berry, 2011). De seguida são apresentadas as definições destes 3 métodos:

***Forward Selection* –** este modelo de seleção inicia-se com um conjunto vazio de variáveis. O melhor atributo original é determinado, segundo um teste estatístico definido, e é adicionado ao conjunto reduzido de variáveis (Han, Kamber, & Pei, 2012). Geralmente, o critério de avaliação baseia-se no calculo do valor de R^2 para cada variável, sendo selecionada aquela com menor valor. Em cada iteração subsequente é adicionado uma nova variável, aquela que, dentro das restantes, acrescentar maior valor ao modelo. Este processo é interrompido quando é atingido um critério de paragem ou quando nenhuma variável melhorar significativamente o modelo ao ser considerada (Linoff & Berry, 2011).

Backward Elimination – o modelo de eliminação de variáveis inicia-se com o conjunto total de variáveis. Em cada iteração, é calculado o R^2 para cada variável, removida aquela com maior valor, e o modelo é novamente reajustado para a próxima iteração (Han, Kamber, & Pei, 2012). O processo

continua até ser atingido um critério de paragem (como o número mínimo de variáveis desejável) ou até nenhuma variável poder ser removida sem perda significativa de valor para o modelo.

Stepwise Selection – este método é a combinação dos dois métodos descritos anteriormente, sendo que, numa primeira instância permanece como um conjunto vazio de variáveis. Em cada iteração do processo pode ser adicionada ou removida uma variável. A flexibilidade deste método acaba por ser vantajosa em cenários em que uma variável introduzida no modelo deixa de ser relevante devido à combinação de efeitos de variáveis subsequentes (Linoff & Berry, 2011).

4.5. Modelos de *Machine Learning*

Esta secção introduz alguns algoritmos de *Supervised Learning* (aprendizagem supervisionada), nomeadamente a regressão logarítmica, árvores de decisão, redes neuronais, *gradient boosting*, e *ensemble.* Como mencionado anteriormente, os objetivos da modelação preditiva são orientados para a compreensão dos dados e descoberta de padrões relevantes, de modo a classificar novos registos de interesse, dado um conjunto de dados previamente classificados (Hand, Mannila, & Smyth, 2001).

4.5.1. Regressão Logarítmica

A regressão logarítmica é um tipo de regressão não linear, tipicamente aplicada quando a variável *target* é binária. O objetivo da regressão logarítmica é estimar a probabilidade de um evento condicional para um conjunto de variáveis (Hosmer & Lemeshow, 1989). Após a estimativa de probabilidade, cada observação é classificada como um evento ou não evento.

Considerando que a variável *target* admite o valor 1 com uma probabilidade *p* de sucesso, e o valor 0 com uma probabilidade de *1-p.* Variáveis com esta natureza seguem uma distribuição de Bernoulli, que é um caso particular da distribuição Binomial quando o número de tentativas é igual a 1. Como a relação entre as variáveis não é linear, é aplicada uma função logarítmica para estabelecer a associação entre as variáveis independentes (*input*) e a variável dependente (*target*).

$$logit(p) = \ln\left(\frac{p}{1-p}\right) \quad (1)$$

No entanto, como a probabilidade *p* é desconhecida, esta tem de ser estimada com base nos valores de *input*. Como resultado, a equação seguinte descreve a relação entre a probabilidade e os respetivos *inputs:* $\bar{\beta}^T X$

$$\ln\left(\frac{p}{1-p}\right) = \bar{\beta}^{T}\mathrm{X} \tag{2}$$

As relações podem ser simplificadas da seguinte forma:

$$\hat{\mathrm{p}} = \frac{1}{1+\mathrm{e}^{-\bar{\beta}^{T}\mathrm{X}}} \tag{3}$$

Os termos do lado direito da equação correspondem a uma função logarítmica. Se for definido $u = \bar{\beta}^{T}\mathrm{X}$, a relação entre a função sigmoide *f* e *u* pode ser observada na figura 4.8.

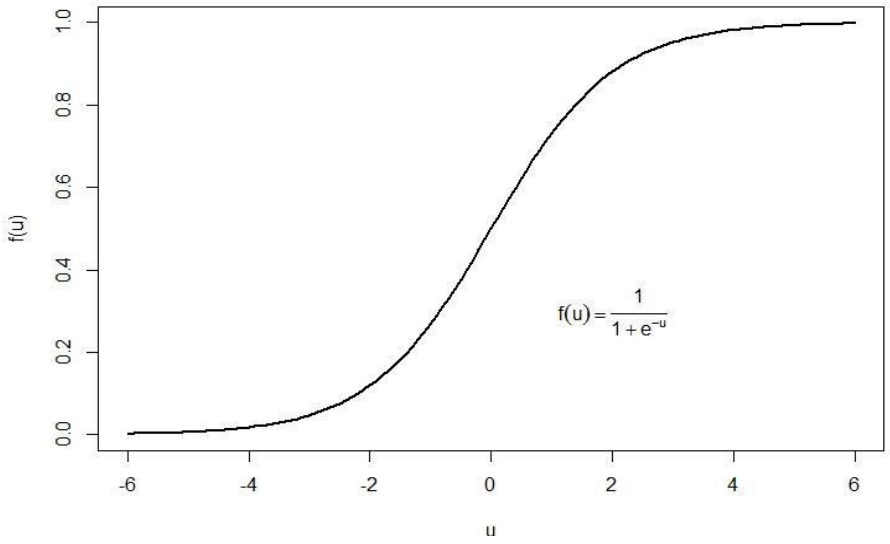

Figura 4.8 - Função Sigmoidal.

Como se pode constatar, para valores elevados de *u* a variável dependente, *f(u)*, aproxima-se de 1, enquanto que para valores muito negativos de *u*, a função aproxima-se de 0. Deste modo, consegue-se obter uma previsão da classificação para cada observação.

4.5.2. Árvores de Decisão

Uma árvore de decisão é uma coleção hierárquica de regras que descreve como dividir um grande conjunto de dados em grupos sucessivamente mais reduzidos. Em cada divisão sucessiva, os membros

dos segmentos resultantes tornam-se cada vez mais semelhantes no que diz respeito à variável *target* (Linoff & Berry, 2011). Uma árvore de decisão é representada por um fluxograma com estrutura em árvore, composta por nós, arcos e folhas. O primeiro nó designa-se por raiz da árvore e representa o atributo mais relevante do conjunto de dados. Cada nó interno representa um teste aplicado a um atributo com o objetivo de o dividir em subconjuntos mais pequenos e homogéneos, enquanto que cada arco representa o resultado do teste, ligando um nó ao nó seguinte ou a uma folha. As folhas representam os nós finas da árvore onde são atribuídas classes aos dados (Han, Kamber, & Pei, 2012). Na figura 4.9 apresenta-se a estrutura base de uma árvore de decisão.

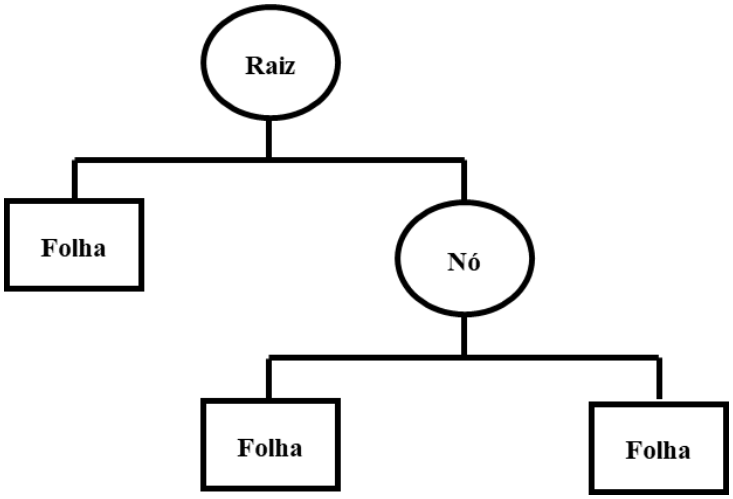

Figura 4.9 - Estrutura base de uma árvore de decisão.

As árvores de decisão constroem conjuntos de regras que são utilizadas como regras de classificação em modelos preditivos. É traçado um conjunto de caminhos desde a raiz até às folhas que permitem testar e classificar facilmente novas observações. A vantagem desta abordagem é que as regras são de fácil compreensão e interpretação, assim como frequentemente revelam processos de negócio não tão evidentes (Hu, 2002).

Shaoling e Yan (2008) abordam este algoritmo aplicado à área de CRM, com o objetivo de descobrir padrões comportamentais nas relações com os clientes, assim como prever e antecipar comportamentos futuros. Segundo os autores, a principal função das árvores é a categorização dos dados e a definição de uma estrutura de regras com base num conjunto de dados conhecidos. Assim, no exemplo demonstrado pelos autores, esta ferramenta é utilizada para aprender a partir de um conjunto de dados de treino, resultando num conjunto de regras designadas por *learning rules* (regras de aprendizagem).

Subjacente à construção da árvore está a seleção de uma medida estatística que avalie a qualidade de partição do nó. Uma das tarefas mais complexas na construção da árvore é precisamente a identificação do atributo com maior poder discriminador (Murthy, 1998). Assim, são propostas pelos autores diferentes abordagens para definir os critérios de partição, sendo as mais frequentes a Entropia e o Coeficiente de Gini. A entropia é uma medida que avalia a incerteza da variável. A função de entropia E de um conjunto de dados S com classificação c é definida como:

$$E(S) = -\Sigma_{i=1}^{c} \mathrm{p}_i * \log_2(\mathrm{p}_i) \quad (4)$$

Na equação, $\mathrm{p_i}$ representa a proporção de S pertencente à classe *i*. Para uma variável *target* binária, a função de entropia pode ser calcula segundo a seguinte fórmula:

$$E(S) = -[\mathrm{p} * \log2(\mathrm{p}) + (1 - \mathrm{p}) * \log2(1 - \mathrm{p})] \quad (5)$$

Na figura 4.10 observa-se a representação gráfica da equação anterior, onde é apresentada a variação da entropia para uma variável binária.

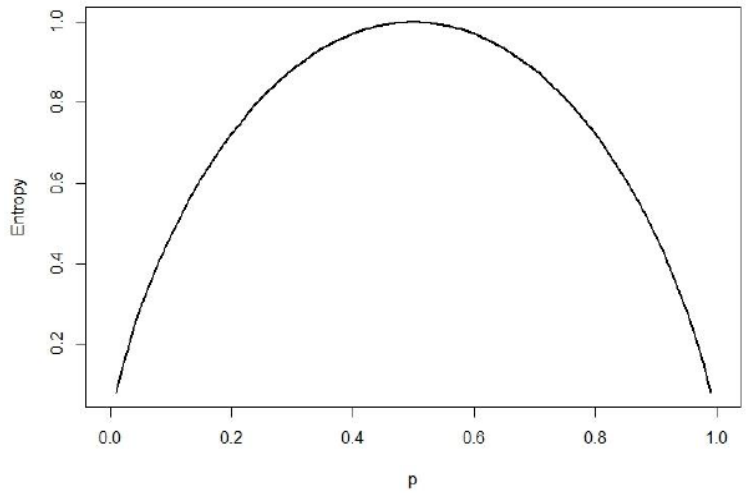

Figura 4.10 - Representação da variação de entropia para uma variável binária.

O valor máximo de entropia é atingido quando a proporção de eventos e não eventos é igual (50%), isto é, quando não há distinção para a variável *target*. O extremo oposto é atingido quando uma das classes detém 100% das observações e a classe complementar 0%. Como o objetivo do algoritmo é encontrar o ponto de partição ótimo que minimiza a entropia e maximiza a diferença de proporção entre as classes da variável *target*, pode falar-se no termo ganho de informação (*information gain* na literatura original) (Du & Zhan, 2002). O ganho de informação mede a redução de entropia causada pela partição dos registos de acordo com uma variável de *input*, tal que:

$$Ganho(S, A) = E(S) - \Sigma_{v \in \text{Valores(A)}} \frac{|S_v|}{|S|} * E(S_v) \quad (6)$$

Onde A corresponde à variável selecionada para proceder à partição de S em S_v subconjuntos. O Ganho(S, A) corresponde à redução de impureza no conjunto S. Assim, a variável que proporcionar o maior ganho é escolhida para efetuar a partição do nó.

A outra medida mencionada como critério de partição, o Coeficiente de Gini. Este coeficiente é calculado tal que, quando todos os indivíduos de um conjunto de dados pertencem à mesma classe, o coeficiente de Gini atinge o seu valor mínimo, igual a 0. Por outro lado, atinge o seu valor máximo, igual a 1, quando os indivíduos estão equitativamente distribuídos pelas classes. Independentemente da medida escolhida, a variável selecionada para a partição do nó será aquela que obtiver um maior ganho de pureza em relação ao nó anterior (Neville, 1999).

Por fim, outro fator que desempenha um papel fundamental na qualidade deste modelo é o tamanho da árvore. Árvores demasiado pequenas poderão não descrever suficientemente bem os dados, enquanto que árvores demasiado grandes correm o risco de aplicar previsões pouco fiáveis. Isto significa que, em certo ponto da construção do modelo, a árvore ajusta-se demasiado ao conjunto de treino, definindo regras para casos particulares como *outliers* ou anomalias nos dados (Du & Zhan, 2002). Para evitar este fenómeno são aplicados critérios de paragem, que permitem parar o crescimento da árvore antes de esta alcançar os estado em que todos os registos são perfeitamente classificados (Neville, 1999), como por exemplo, o número mínimo de indivíduos por folha, o número mínimo de indivíduos por nó, a profundidade máxima da árvore, ou o número máximo de divisões por nó.

De modo a evitar a sobreaprendizagem dos dados, pode ainda recorrer-se ao desbaste da árvore (*Tree Pruning* na literatura original), método que procura identificar e remover folhas e nós responsáveis pela partição em conjuntos muito específico. Combinados, estes dois procedimentos aumentam o poder generalizador deste classificador, garantindo a precisão da classificação em dados desconhecidos (Neville, 1999).

Assim, pode afirmar-se que as árvores decisão são dos modelos mais utilizados no *data mining*, devido à versatilidade de problemas a que pode ser aplicado, e à interpretabilidade dos resultados produzidos (Linoff & Berry, 2011). Destacam-se também por aceitarem como *input* vários tipos de variáveis (nominais, ordinais e intervalares) e lidarem com *missing values*.

4.5.3. Redes Neuronais

As Redes Neuronais (*Neural Networls* na literatura original) são um conjunto de técnicas "poderosas, flexíveis e robustas", utilizadas para resolver problemas complexos, onde se procura estimar ou classificar resultados futuros (Linoff & Berry, 2011). Esta designação tem origem na semelhança estrutural e funcional com o cérebro humano. Para muitos autores, a aplicação de redes neuronais a questões reais é bastante atrativa, dada a sua capacidade de lidar com problemas não lineares e anomalias dos dados. (Basheer & Hajmeer, 2000).

Numa rede neuronal, o comportamento base de um neurónio consiste na receção de *inputs*, na transformação desses valores aplicando uma função de ativação, e respetiva produção de um *output*. Na literatura existem vários tipos de redes neuronais, distinguindo-se pelas diferentes regras de aprendizagem e topologias. As arquiteturas mais conhecidas são a do perceptrão e a do perceptrão multicamadas (também conhecido como *Multi-layer perceptron, MLP*, na literatura original) (Linoff & Berry, 2011).

O perceptrão é a rede neuronal mais simples que existe, com apenas um neurónio, utilizada apenas para classificações linearmente separáveis (Basheer & Hajmeer, 2000; Han, Kamber, & Pei, 2012). Historicamente, o modelo do perceptrão multicamadas surge como resolução para problemas não lineares. A diferença estrutural de ambos os modelos reside na existência de camadas intermédias, também conhecidas como camadas escondidas (ou *hidden layers* na literatura original) (Basheer & Hajmeer, 2000). Na figura 4.11 ilustra-se a estrutura típica de uma rede neuronal.

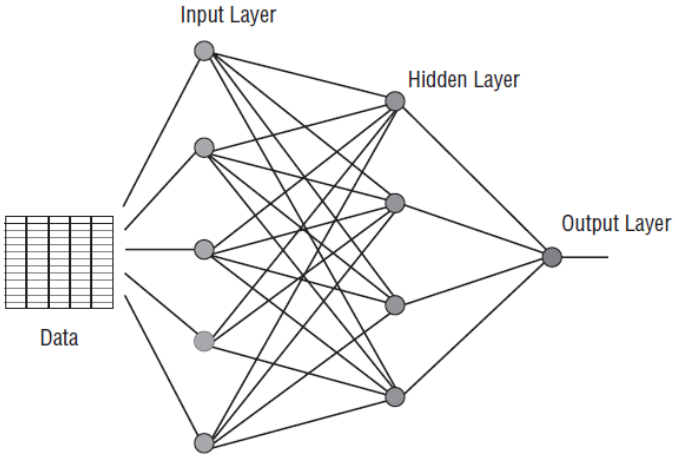

Figura 4.11 - Estrutura básica de uma rede neuronal multicamadas (MLP). Adaptado de Linoff & Berry (2011).

Na representação acima, podem observar-se três camadas: a camada de entrada (*input layer*), que é o ponto de entrada dos dados na rede; a segunda camada conhecida como camada escondida (*hidden layer*), onde os neurónios recebem o *output* da camada de entrada; e a camada de saída (*output layer*), que combina os resultados da camada anterior e produz o resultado da rede. As ligações entre as várias camadas representam os pesos, que são os fatores de multiplicação dos respetivos *inputs* (Basheer & Hajmeer, 2000).

Apesar do modelo de funcionamento deste algoritmo ser simples, os resultados produzidos em conjuntos de dados complexos tornam-se de difícil interpretação e avaliação crítica (Hu, 2005). Numa primeira fase dá-se a propagação para a frente dos valores, começando pela camada de entrada. Os valores produzidos são transmitidos para os nós da camada escondida. O resultado de cada nó dessa camada é calculado como a combinação linear dos seus *inputs* e respetivos pesos. Posteriormente, os valores resultantes são aplicados a uma função de ativação e propagados até à camada de saída. A função de ativação tipicamente utilizada é a função sigmoide, dadas as propriedades já anteriormente mencionadas (Basheer & Hajmeer, 2000). Como descrevem Linoff e Berry (2011), uma rede neuronal pode ter um número infinito de camadas escondidas, contudo, uma única é suficiente para a maioria dos problemas. Os autores advertem que, embora os neurónios contribuam para aprendizagem dos dados, quanto maior o número de unidades na camada escondida, maior a capacidade de memorização, aumentando o risco de *overfitting*. Basheer e Hajmeer (2000) referem como os maiores desafios na aplicação deste método, a determinação do número ótimo de neurónios na camada escondida, assim como a atribuição de pesos a cada ligação.

Determinados os resultados obtidos da rede, estes são comparados com os resultados esperados, e é calculado o erro associado. Esta é a fase de aprendizagem do algoritmo, onde ocorre o fenómeno de retropropagação (*backpropagation* na literatura original). Esta fase do algoritmo tem como objetivo produzir um resultado com o menor erro possível, recorrendo, para tal, ao ajuste dos pesos atribuidos a cada ligação. Inicia-se assim, a retropropagação dos resultados para a camada de saída e para as camadas escondidas. Em cada iteração do algoritmo, os pesos são ajustados em função da diferença entre o resultado esperado e o resultado obtido, procedendo no sentido da camada de saída para a camada de entrada. Esta metodologia identifica os pesos que têm uma maior contribuição para o erro, e atualiza-os de modo a obter melhores resultados. Na parametrização deste algoritmo é também importante referir a taxa de aprendizagem, um valor constante que define a proporção de atualização atribuída a cada peso. Taxas de aprendizagem demasiado pequenas tornam mais moroso o processamento do algoritmo, enquanto que taxas demasiado elevadas potenciam a instabilidade do modelo, produzindo resultados divergentes.

4.5.4. Métodos de *Ensemble*

Os métodos de *ensemble*, como o próprio nome indica, consiste, na combinação de dois ou mais modelos (Dean, 2014) para criar um novo modelo capaz de classificar novas instâncias. Ao combinar vários classificadores e respetivos resultados, muitas vezes produzem-se previsões mais precisas do que considerando os algoritmos individualmente, o que confere um maior poder preditivo ao modelo final (Dean, 2014; Kumar, Kongara, & Ramachandra, 2013; Hu, 2005). Uma vez que pode ser treinado e utilizado para fazer previsões, o próprio *ensemble* (combinação) pode ser considerado um algoritmo de aprendizagem supervisionada (*supervised learning)*.

Enquanto que um classificador utiliza um único conjunto de regras para classificar um evento futuro, um classificador *ensemble* aplica uma combinação de regras aprendidas pelos modelos originais para tomar as decisões de classificação. Esta abordagem tem a vantagem de atenuar o peso dos erros existentes em cada modelo, sendo mais eficaz na sua previsão final. Uma vez que incorpora o resultado de vários modelos numa única previsão, considera-se também que tem um maior poder de generalização, contornando os fenómenos de *overfitting.* Assim, considera-se que teoricamente, este método tem maior capacidade preditiva perante uma nova instância (Hu, 2001).

No entanto, é fundamental realçar que os métodos de *ensemble* só conseguem ter um desempenho substancialmente melhor do que os restantes modelos quando os modelos considerados são complementares, isto é, quando discordam na classificação de eventos. Se não houver variabilidade entre os *inputs* do método de *ensemble*, acrescentando todos a mesma informação, não são esperadas melhorias nos resultados. Vários autores referem-se a esta propriedade, combinando diferentes modelos como redes neuronais, árvores de decisão e regressões no desenho dos seus modelos. Kumar, et. al (2013), por exemplo, mencionam que estas técnicas tendem a produzir melhores resultados quando existe uma diversidade significante entre os modelos, e portanto, propõem a combinação de algoritmos aleatórios, como árvores de decisão, para produzir modelos mais robustos e poderosos. Na literatura podem encontrar-se vários métodos de *ensemble* (Dean, 2014; Hu, 2001; Kumar, Kongara, & Ramachandra, 2013; Linoff & Berry, 2011; Han, Kamber, & Pei, 2012), dos quais os mais populares:

- *Bagging*
- *Gradient Boosting*
- *Random Forest*
- *Random Trees*
- *Alternating Decision Trees*

Os vários métodos têm alguns aspetos em comum, dos quais as alternativas de combinação, isto é, as funções que definem como é que os resultados de cada modelo combinam. Face a este tema, podem ser adotadas três estratégias distintas, das quais: a média, a votação e o valor máximo.

O valor médio é essencialmente aplicado em modelos de regressão, pois tira partido dos valores contínuos das probabilidades calculadas para cada observação. O método de votação é tipicamente aplicado em problemas de classificação com variáveis *target* categóricas, onde cada modelo vota no resultado esperado de cada instância, e o resultado da classificação é aquele que receber a maioria dos votos. Quando se opta pelo máximo, é considerado o valor máximo das probabilidades previstas, em *targets* intervalares, ou o valor mais frequente, em *targets* categóricos.

Na secção seguinte, é descrito com maior detalhe o método de *Gradient Boosting*, um caso particular dos métodos *ensemble* exclusivamente com àrvores de decisão.

4.5.4.1. Gradient Boosting

O algoritmo de *machine learning Gradient Boosting* é um caso particular de um método de *ensemble*, no sentido em que resulta do encadeamento sucessivo de modelos preditivos. Tipicamente, os classificadores utilizados na construção deste algoritmo são árvores de decisão. O conceito deste modelo baseia-se no treino em série dos classificadores, permitindo que cada classificador aprenda com os erros do classificador anterior, concentrando o esforço dos classificadores subsequentes nas instâncias mais difíceis de aprender, minimizando incrementalmente o erro total do algoritmo, e melhorando assim a precisão da previsão (Dean, 2014).

Kumar, et al., (2013) na sua obra descreve este modelo como uma técnica de classificação que consiste na criação de um classificador forte através da combinação de vários classificadores mais fracos. Dean (2014) vai mais longe, explicando como decorre o processo. Basicamente, o conjunto de treino utilizado por cada classificador da série é selecionado com base no desempenho do classificador anterior. A tendência será escolher instâncias que tenham sido mal classificadas anteriormente, proporcionando novos classificadores com maior capacidade de classificação. A seleção dos registos assenta num mecanismo de ajustamento do peso de cada registo em cada iteração, sendo que os registos com maior erro têm maior probabilidade de serem considerado no classificador seguinte. O processo termina quando é atingido um dos critérios de paragem definidos, que poderá ser o número de modelos produzidos ou até mesmo a precisão mínima aceitável.

Teoricamente, o método de *Gradient Boosting* reduz a influência de valores extremos e é menos suscetível à sobreaprendizagem dos dados, comparativamente a uma única árvore de decisão. Porém, tal como uma árvore de decisão, o modelo de *boosting* não faz suposições sobre a distribuição dos dados, dependendo apenas do processo contínuo de aprendizagem.

4.6. Métodos e Métricas de Avaliação

A fase de avaliação desempenha um papel importante quando aplicado a algoritmos de *machine learning*, uma vez que permite inspecionar a qualidade dos resultados produzidos. Nesta secção, apresentam-se um conjunto de métodos e métricas que devem ser consideradas para determinar a confiança nos resultados e previsões finais.

Existe um vasto conjunto de métricas que permitem testar efetivamente a qualidade dos dados produzidos. No que respeita a modelos de classificação para variáveis dependentes discretas, o método mais aplicado é a Matriz de Confusão, demonstrado na tabela 4.1. A Matriz de Confusão é uma representação do conjunto de dados em análise, onde as linhas exibem as classes reais e as colunas as classes previstas. Em classificações binárias, as instâncias podem ser rotuladas como positivas ou negativas, correspondendo a 1 ou 0, respetivamente (Kohavi & Provost, 1998). Um classificador eficaz é capaz de distribuir a maior parte dos registos na diagonal da matriz (previsões verdadeiras), com as restantes entradas sendo próximas de zero (Han & Kamber, 2001).

Tabela 4.1 - Matriz de Confusão em problemas de Classificação Binária.Adaptado de Kohavi & Provost, 1998.

		Classe Prevista	
		Positivo	**Negativo**
Classe Real	**Positivo**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Negativo	Falso Positivo (FP)	Verdadeira Negativo (VN)

Na matriz, os verdadeiros positivos (VP) denotam os eventos positivos corretamente classificados, da mesma forma que os verdadeiros negativos (VN) exibem o número de eventos negativos corretamente classificados. Falsos Positivos (FP) e Falsos Negativos (FN) representam os eventos positivos e negativos, respetivamente, erradamente classificados. Relativamente às métricas, destacam-se as seguintes (Davis & Goadrich, 2006; Kohavi & Provost, 1998):

- **Sensibilidade** (*Sensitivity*) – Também conhecida como Taxa de Verdadeiros Positivos, esta taxa mede a proporção de registos positivos que são corretamente classificados.

$$Sensibilidade = \frac{VP}{VP + FN} \tag{7}$$

- **Especificidade** (*Specificity)* – Também conhecida como Taxa de Verdadeiros Negativos, esta taxa mede a proporção de registos negativos que são corretamente classificados.

$$Especificidade = \frac{VN}{VN + FP} \tag{8}$$

- **Precisão** (*Precision)* – Esta taxa representa a proporção de registos realmente positivos no conjunto de registos previstos como positivos.

$$Precisão = \frac{VP}{VP + FP} \tag{9}$$

- **Exatidão** (*Accuracy)* – Esta taxa representa a proporção de registos corretamente classificados no total de todas as classificações.

$$Exatidão = \frac{VP + VN}{VP + VN + FP + FN} \tag{10}$$

- **Curva ROC e AUC** – siglas para *Receiver Operating Characteristic Curve* e *Area under the Curve.* Esta métrica é uma representação gráfica que exibe a variação da taxa de verdadeiros positivos com a taxa de falsos negativos. No eixo das abcissas é apresentada a proporção de registos negativos mal classificados, $\frac{FP}{VN+FP} = 1 - Especificidade$, e no eixo das ordenadas a proporção de registos positivos que são corretamente classificados (Sensibilidade).

  Posteriormente é calculada a área debaixo da curva ROC, sendo que o melhor modelo é aquele com maior área, ou seja, com o valor estatístico mais próximo de 1 (Hand, Mannila, & Smyth, 2001).

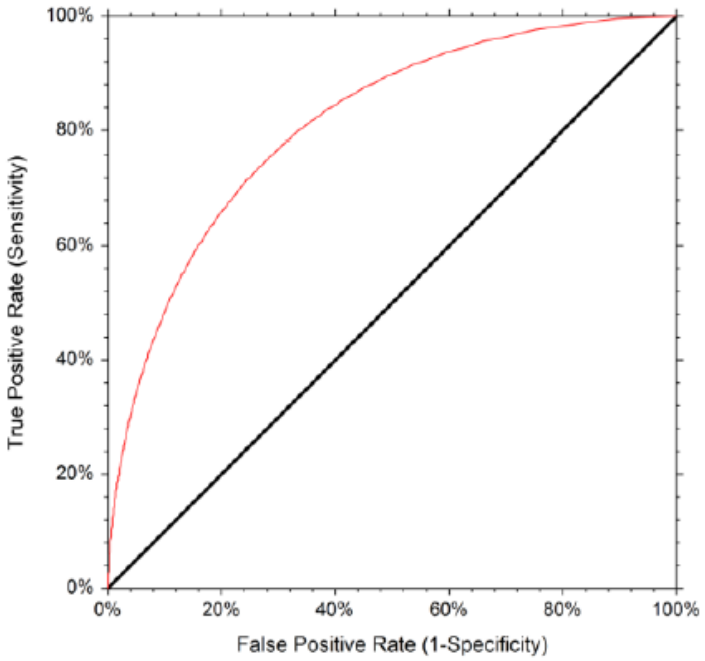

Figura 4.12 - Representação da Curva ROC.

Apesar das métricas apresentadas serem as mais recorrentes, não existe um consenso relativamente a qual o melhor conjunto. A aplicabilidade e relevância de cada métrica está altamente relacionada com o problema em causa.

5. Metodologia para otimização do desempenho da gestão de *Leads*

Nos capítulos anteriores refletiu-se sobre a importância e o papel das *leads* nas estratégias de aquisição de novos clientes nas empresas. De facto, demonstrou-se que este tema começou a ser pertinente há cerca de uma década (Metzger, 2005; Goldie, 2007), mas que atualmente continua a ser considerada uma abordagem disruptiva e inovadora (Gordon, 2018; Krozel, 2019). Porém, e apesar do seu potencial, é ainda reconhecido por muitas empresas como um conjunto de processos bastante ineficientes, arbitrários, e com baixos resultados na conversão (Monat, 2011; Gordon, 2018).

Embora existam na literatura vários estudos que demonstram os bons resultados conseguidos da aplicação de técnicas de *data mining* em problemas relacionados com a gestão de clientes (Ling & Li, 1998; Greenyer, 2000; Zhang, Yang, Shi, & Lu, 2008; Moro, Laureano, & Cortez, 2011), não foi ainda conduzida nenhuma investigação que integre estas metodologias na gestão de *leads*. A presente investigação tem como objetivo propor uma metodologia generalizada de gestão de *leads*, que permita tomar decisões baseadas e suportadas em dados e factos reais. Os capítulos anteriores foram essenciais para reunir uma base sólida e forte de fundamentação teórica acerca do que é pretendido em cada uma das etapas da metodologia proposta. No entanto, tal como afirma Compton (2012), no desenho de qualquer metodologia ou processo devem ser consideradas as especificidades e necessidades de funcionamento de cada negócio, resultando numa solução customizada e à medida de cada empresa.

Esta metodologia compreende um paralelismo entre o que é o ciclo normal de funcionamento da gestão de *leads* com um conjunto de ações e medidas desempenhadas do lado do negócio. O que se pretende demonstrar é que as atividades de negócio devem acompanhar o fluxo de vida das *leads*, e que uma aplicação oportuna destas medidas tem um impacto positivo no desempenho de cada processo, potenciando o sucesso das conversões finais. Portanto, esta metodologia assenta na premissa de que as *leads* podem ter diferentes níveis de qualificação e que cada nível requer ações de negócio específicas, sendo que aquelas com maior qualidade são convertidas em clientes com maior recorrência (Silverstein, 2012).

Na figura 5.1 apresenta-se a metodologia proposta, resumida em forma de diagrama. Do lado esquerdo encontram-se os processos referentes à gestão de *leads*, conforme descrito no capítulo 3.2.1. No lado direito do diagrama, representou-se o fluxo de atividades e ações de negócio que devem ser executadas paralelamente, e que têm um impacto direto nas etapas da gestão de *leads.*

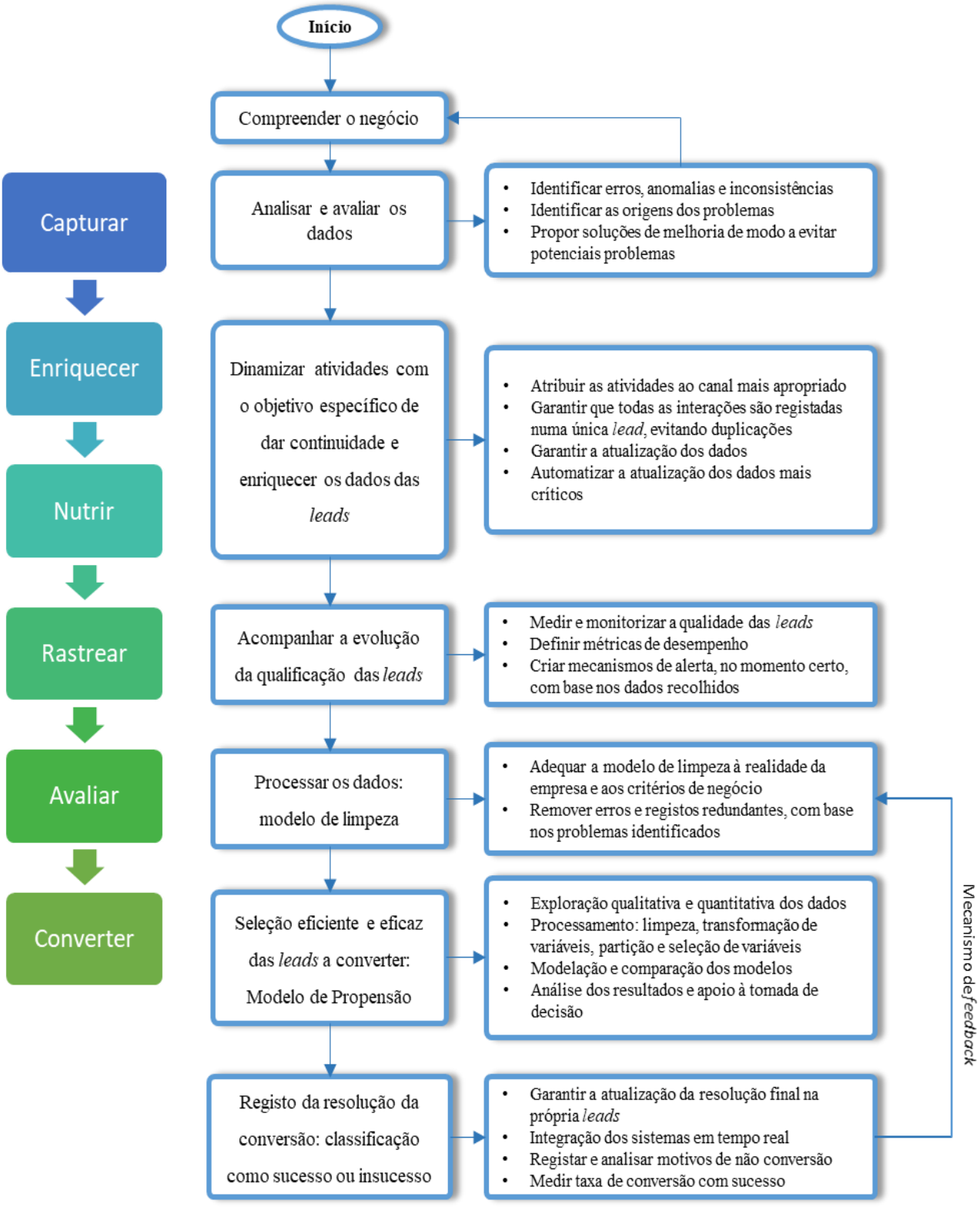

Figura 5.1 – Diagrama da metodologia proposta.

Compreensão do Negócio

Conforme representado no diagrama, a compreensão do negócio é a fase inicial e da qual derivam todas as restantes análises, conclusões e ações aplicadas. Tal como o nome indica, esta fase inclui a compreensão dos objetivos da empresa e do projeto em questão, assim como a dos requisitos de negócio. O objetivo final será sempre a conversão das *leads* em clientes, contudo, compreender o modelo de funcionamento dos canais de venda, como é processada a própria venda e as respetivas interações entre canais, são exemplos de noções essenciais para a correta configuração de todo o processo (Samuels, 2013).

No entanto, no ponto de vista desta metodologia, esta fase envolve também a análise dos dados pretendidos, isto é, uma previsão de qual será a informação útil, tanto na fase de qualificação como na fase de conversão. Definido o tipo de informação que se pretende, deve materializar-se o conjunto de dados que será necessário (Chapman, et al., 2000). Ou seja, o objetivo é que seja determinado desde o início o tipo de dados e de informação que se pretende, para que exista uma avaliação dos dados existentes e para que se encontre a melhor solução para a captura de novos dados. Como referiu Olson (2003) no seu trabalho, seria vantajoso para esta etapa envolver também as áreas técnicas, uma vez que detêm maior conhecimento dos sistemas de informação internos de cada empresa.

Análise e Avaliação dos dados

Posteriormente, e como representado na figura 5.1, segue-se a fase de análise e avaliação do histórico de dados capturados. Esta fase serve como uma auditoria às fontes de dados, de modo a perceber se estão a ser cumpridos os requisitos definidos na fase anterior. Atualmente, a dinâmica das empresas tem evoluído no sentido de promover e estimular a captura de *leads* em todas as interações com o cliente, envolvendo os vários canais de comunicação (Willis & Flo, 2016). Com o crescimento do número de *leads* geradas nos canais *online* (Bairstow, 2016), o volume, a variabilidade e a heterogeneidade dos dados tem tornado mais complexo o seu tratamento e a análise (Silverstein, 2012). Na sua investigação, Krozel (2019) afirma que algumas das principais causas dos problemas estão também associadas à desatualização e incompatibilidade das ferramentas. Por sua vez, os problemas enumerados focam a fraca qualidade, completude e veracidade dos dados, reforçando a existência de duplicação de registos e falta de normalização das variáveis (Krozel, 2019).

Portanto, o objetivo desta fase passa por analisar com sentido crítico os dados recolhidos, e à luz dos processos de captura implementados. Só assim é possível identificar erros, anomalias e inconsistências

existentes nos dados (Silverstein, 2012; Compton, 2012). Após a deteção dos problemas existe, idealmente, um mecanismo de correção que permite corrigir e alterar estes problemas diretamente na fonte, de modo a que não se propagarem em dados futuros (Olson J. E., 2003). Este mecanismo dependerá sempre do problema e do caso em estudo, mas pode passar por envolver um requerimento técnico de alterações na implementação, ou simplesmente um reforço dos procedimentos junto de canais assistidos.

Atividades de qualificação e nutrição

Posto isto, é necessário trabalhar sobre as *leads* capturadas. Como já foi referido anteriormente, a qualificação e acompanhamento das *leads* é fundamental para que exista uma conversão final. Conforme admite Krozel (2019), este tipo de atividades tem um impacto positivo na taxa de sucesso e na eficiência de todo o processo. Embora conceptualmente as fases de enriquecimento e nutrição tenham objetivos e pressupostos diferentes, do ponto de vista de negócio não há uma separação linear entre ambas. Como se pode observar na figura 5.1, o modelo sugere que as atividades de dinamização da qualificação e da nutrição de *leads* ocorram simultaneamente. De facto, o que se verifica no contexto de uma empresa é que todas as interações com uma *lead* são oportunidades de enriquecimento, em termos de dados, e de nutrição, estimulando a relação e o interesse da *lead* (Gillin & Schwartzman, 2011).Só a coordenação entre estas duas atividades permite melhorar efetivamente a qualidade dos dados, melhorando assim o entendimento sobre a *lead* e as respetivas ações de seguimento que devem ser tidas em consideração.

Assim, a dinamização das atividades de qualificação e de nutrição podem requerer várias iterações, e como tal, devem acompanhar os desenvolvimentos e garantir que existe um registo preciso dos novos dados fornecidos (Silverstein, 2012). Para além disso, e resultado da avaliação dos dados e das conclusões fornecidas por várias empresas (Velocify, 2012; Krozel, 2019), estas atividade devem também promover a utilização de uma única *lead* em todas as iterações, de modo a evitar duplicações de registos. Como se constatou, pelo facto da conversão na maior parte dos casos não ser instantânea, são necessárias interações posteriores com as *leads*, que resultam na criação de novas *leads* com a mesma informação das anteriores. Assim, esta metodologia sugere que seja garantida a atualização das *leads* em todas as interações, quer por mecanismos automáticos, que manualmente. Idealmente, a automatização de processos repetitivos e que podem ser robotizados é sempre preferível, uma vez que é menos falível e produz melhores resultados.

Um caso bastante comum, partilhado por Silverstein (2012), é a atribuição de *leads* geradas em canais *online* a canais de vendas comissionados. Um dos problemas levantado pelo autor é a falta de motivação ou recompensação dos assistentes de vendas pelas atividades de qualificação de *leads*¸ uma vez que estes são comissionados pelo número de vendas efetivas que realizam. Portanto, parte das atividades sugeridas na metodologia devem também incluir a definição de objetivos específicos e direcionados para a qualificação da informação recolhida junto dos canais de vendas.

Rastrear

Parte da estratégia de gestão de *leads* passa por comunicar com a *lead* certa, no momento certo, e com a abordagem e oferta mais adequadas. A fase de qualificação anterior permite estabelecer qual a melhor abordagem a adotar, ou qual a oferta mais tentadora para determinada *lead.* Por sua vez, a fase de rastreamento compreende o acompanhamento da evolução de uma *lead*, através de métricas pré-definidas que permitem medir o desempenho de cada iteração e aptidão da *lead* para ser contactada novamente (Gordon, 2018). Sendo esta uma atividade constante e transversal a toda a fase de qualificação, idealmente seria automatizada e integrada com as ferramentas utilizadas. Uma das muitas vantagens da automatização passa pela possibilidade de criar mecanismos de alerta que, com base na informação contida nos dados, identifica e notifica os utilizadores do momento certo para contactar uma *lead*, melhorando por si só a eficiência da seleção de *leads.*

Limpeza dos dados

Na metodologia apresentada, a fase de avaliação de *leads* corresponde à construção de um modelo de limpeza de dados. Debateu-se anteriormente que as bases de dados são altamente suscetíveis a inconsistências, falhas e ruídos nos dados. Segundo Han, Kamber e Pei (2012) isto deve-se maioritariamente à diversidade heterogénea de fontes de dados, assim como, à dimensão e volume de dados armazenados. Em muitas empresas, existe também a particularidade de muitos dados serem atualizados ou preenchidos manualmente, adicionando variabilidade aos valores recolhidos e potenciando a ocorrência de erro (Zhao, Wu, & Gao, 2008). Empiricamente é sabido que dados com pouca qualidade produzem resultados com pouca qualidade, portanto, são vários os autores que evidenciam a importância do processamento prévio dos dados, de modo a reduzir o efeito dos erros e ruídos nas fases seguintes (Han, Kamber, & Pei, 2012), assim como a complexidade e os custos de processamento (Rahm & Hai Do, 2000).

Como foi abordado no capítulo 4.4.1, as rotinas de tratamento de dados tendem a resolver três dos principais problemas, sendo eles *missing values, outliers* e erros, no entanto, as abordagens de resolução devem ser flexíveis e adequadas à realidade dos dados e do negócio em estudo (Han, Kamber, & Pei, 2012). Deste modo, a metodologia sugere que sejam considerados os erros e os problemas encontrados na fase de avaliação inicial, assim como as regras de negócio resultantes da compreensão realizada, e estes sejam traduzidos em regras analíticas. A coordenação destas regras constitui um modelo de limpeza, que idealmente permita excluir registos errados ou redundantes, transformar variáveis resolvendo inconsistências ou facilitando a sua interpretação, entre outras operações. Embora esta seja uma fase mais morosa do projeto, a correta limpeza, integração, tratamento e transformação dos dados é essencial para obter um modelo preciso na fase seguinte (Hu, 2005).

Segmentação eficaz das *leads*

Por fim, segue-se a fase de conversão que corresponde ao verdadeiro objetivo das empresas e de todo o processo de gestão de *leads.* Segundo a metodologia apresentada, esta fase deve ser assistida por um conjunto de ações de negócio com a finalidade de tornar mais eficiente e eficaz a seleção de *leads.* Assim, a proposta inclui um modelo de propensão capaz de auxiliar a tomada de decisão no processo de segmentação, e, consequentemente aumentar a taxa de conversão.

Através da revisão de literatura realizada no âmbito da aplicação de técnicas de *data mining* em problemas reais de gestão de clientes, detetou-se vários autores que propõe a construção de modelos analíticos como suporte à decisão (Hu, 2005; Chitra & Subashini, 2013; D'Haen & Poel, 2013; Lu, Lin, Lu, & Zhang, 2014; Gupta, Wasid, & Ali, 2016). Este tipo de modelos permite aprender e extrair padrões de comportamento do histórico dos dados, classificando posteriormente novos registos com base no conhecimento adquirido (Linoff & Berry, 2011). Aplicados à gestão de *leads*, estes métodos podem ser utilizados para estimar a propensão de conversão de cada *lead*, e assim, gerir de forma mais eficiente os recursos e garantir uma maior taxa de sucesso.

Deste modo, a metodologia sugere a construção de um modelo de propensão com base nos sucessos e insucessos das conversões anteriores, com o objetivo de estimar a probabilidade de conversão de cada *lead* aberta na base de dados. Um modelo de propensão é um modelo de classificação preditivo, tal como foi descrito no capítulo 4, referente ao *Data Mining.* A construção de um modelo preditivo é composta por uma sequência de vários procedimentos que garantem a qualidade e confiança no resultado final, como por exemplo: a exploração qualitativa e quantitativa dos dados de treino; o processamento dos dados através de ações de limpeza, transformação, partição e seleção de variáveis; a aprendizagem

segundo vários modelos; a avaliação e comparação estatística dos modelos; e a análise final dos resultados como apoio à tomada de decisão. Todas estas técnicas e procedimentos foram abordados em detalhe ao longo do capítulo 4, com destaque da seleção de variáveis, modelação e seleção de modelos, que são fases que exigem um maior conhecimento estatístico e matemático. Assim, como resultado do modelo é esperado um mecanismo automático que estime a propensão de um conjunto de *leads*, permitindo a identificação e seleção daquelas com maior probabilidade de conversão.

Conversão

Por último, a metodologia compreende a fase de conversão de uma *lead* que, apesar dos esforços depositados, pode resultar em sucesso ou insucesso. Em capítulos anteriores referiu-se que a competência e a eficácia dos canais *offline* têm uma forte influência nesta fase do processo (Bairstow, 2016). Tal como foi demonstrado pelas publicações de diversos autores, tipicamente recorre-se a canais de *telemarketing* ou de vendas diretas na fase de conversão final (Silverstein, 2012). Nesse sentido, é necessário formar as equipas com os procedimentos corretos e orientá-las para as iniciativas que promovem a qualidade das *leads*. Desse modo, a metodologia propõe que as atividades de negócio referentes a esta etapa promovam e garantam, em primeiro lugar, o registo da informação resultante de cada tentativa de conversão, e em segundo, a atualização da própria *lead*. Este reforço de procedimentos ganha importância uma vez que os principais problemas reportados incidem precisamente na falta de qualificação e duplicação dos dados. Idealmente, existiria uma classificação automática, de sucesso ou insucesso, no final de cada iteração. Esta integração permitiria reduzir os erros associados à introdução manual, para além de que, todas as atualizações estariam disponíveis, em tempo real, para análise e para melhoria do modelo.

Observando o diagrama ilustrado na figura 5.1, pode constatar-se que não existe um fim no fluxo de atividades. Isto porque, segundo a metodologia, o processo de gestão de *leads* e as iniciativas de conversão devem ser contínuas e autossustentáveis ao longo do tempo. Embora esta não seja uma realidade para muitas empresas, idealmente, no final da conversão existe um mecanismo de retroação que incorpora a informação resultante de cada conversão de novo no fluxo de dados. A introdução destes dados enriquece a informação de treino do modelo, afinando e tornando o modelo mais preciso a cada iteração. A este mecanismo chama-se mecanismo de *feedback*, pois funciona como uma resposta à estimativa prevista inicialmente pelo modelo (Silverstein, 2012). Ao comparar o resultado previsto com o resultado real, é possível medir a precisão real do modelo e ajustar a aprendizagem em concordância. Portanto, para que exista um ciclo orgânico de alimentação dos dados de treino (classificados), é fundamental que no final de cada conversão ocorra o registo da variável dependente, como sucesso ou insucesso.

Como forma de medição do desempenho e avaliação das atividades desenvolvidas, devem ser definidas métricas estratégicas que possibilitem a monitorização da taxa de conversão. Este tipo de métricas deve ser adaptado à realidade do negócio, incidindo no desempenho de campanhas, canais de vendas, processos de segmentação, entre outros. Só ao comparar taxas de sucesso em eventos sucessivos, ou avaliando a sua oscilação, é possível medir o impacto real que este tipo de metodologias tem nas empresas (Samuels, 2013).

Por fim, e embora esta fase incida nos eventos de conversão, a metodologia também dá destaque aos eventos de não conversão, recomendando que deve existir o registo contínuo dos motivos de não conversão. Da mesma forma que uma previsão correta reforça o modelo, uma precisão incorreta também acrescenta valor. Alguns autores mencionam inclusive, que é mais interessante para um modelo receber resultados de previsões incorretas, uma vez eu vão adicionar novo conhecimento ao modelo, permitindo que a aprendizagem se reflita em previsões futuras (Lu, Lin, Lu, & Zhang, 2014). Assim o registo dos motivos de não conversão demonstra ser bastante útil, pois permite que os dados sejam analisados futuramente e transformados em informação útil e valiosa para o negócio.

6. Caso de Estudo – Gestão de *Leads* numa empresa de telecomunicações

Neste capítulo é apresentado o caso de estudo no qual se fundamentaram os desenvolvimentos da investigação. Como resultado da análise, foram aplicadas várias técnicas e metodologias que permitiram resolver e atingir os objetivos propostos inicialmente.

6.1. Metodologia

Este capítulo engloba a apresentação e compreensão do caso de estudo. Atendendo ao objetivo final de aumentar a taxa de conversão em *leads* e produzir uma ferramenta inteligente que suporte a tomada de decisão no processo de segmentação de *leads*, foi necessário efetuar uma análise exaustiva da implementação atual e do modelo de funcionamento na empresa. Este foi o ponto de partida para todos os desenvolvimentos seguintes.

Primeiramente, procurou-se identificar as principais dificuldades e obstáculos, o que conduziu à deteção de um grave problema na qualidade dos dados registados. Deste modo, avançou-se com o desenvolvimento de um modelo de limpeza, construído através um conjunto de regras de negócio. O pré-processamento dos dados foi uma etapa essencial para garantir a melhoria significativa dos dados que seriam utilizados para a construção do modelo preditivo, assegurando maior qualidade e confiança nos resultados produzidos.

Porém, o maior esforço foi aplicado no desenvolvimento do modelo de propensão de *leads*. Como mencionado anteriormente, nos processos de modelação preditiva, incorporados no processo de Descoberta de Conhecimento em Base de Dados, podem ser adotadas diferentes abordagens das quais CRISP-DM e SEMMA. A metodologia adotada na componente prática da presente dissertação resultou da combinação de ambas as estratégias. Enquanto que a metodologia CRISP-DM participa com a fase de compreensão do negócio, onde é realizada a caracterização do negócio e especificados os requisitos técnicos e estratégicos, a metodologia SEMMA serve a componente funcional de construção do modelo, preconizando a seleção e compreensão dos dados, exploração, transformações, modelação e avaliação dos resultados. Note-se que a metodologia SEMMA foi totalmente construída recorrendo à ferramenta *SAS Enterprise Miner*.

Por fim, foram apresentados os resultados e debatidas as possíveis soluções a adotar com base nas propensões calculadas.

6.2. Caracterização do Caso de Estudo

O caso de estudo apresentado tem por base a realidade de uma empresa de telecomunicações. Como já foi referido, o sector das telecomunicações é altamente competitivo, e, portanto, as empresas procuram destacar-se através da customização da oferta, inovação e qualidade dos serviços.

Na última década, a grande disrupção deste setor em Portugal ocorreu ao nível da fibra ótica e do lançamento do serviço de casa. Enquanto que anteriormente o foco era exclusivo no serviço móvel, mais recentemente o desafio foi o crescimento do serviço fixo. Este serviço tem algumas particularidades que tornam o processo de adesão mais complexo: o cliente não pode estar fidelizado a outra operadora, tem de existir instalação de fibra ótica até à casa do cliente e envolve a instalação de equipamentos na casa do cliente. Este são fatores que condicionam a adesão ou alteração do serviço.

Como tal, esta empresa adotou um modelo de gestão de *leads* como estratégia de aquisição de clientes para o serviço fixo. O objetivo passava por capturar *leads* em todas as interações com possíveis clientes que demonstrassem interesse, instituindo esses procedimentos nos vários canais. Assim, existem 3 fontes principais de criação de *leads*, sendo eles os parceiros de telemarketing, os assistentes nas lojas físicas e o website, onde os utilizadores são convidados a inserir os seus dados para verificar a cobertura na sua morada. O canal dominante na produção de *leads* e que consegue chegar a um maior número de consumidores é o *online*, contudo, é também o mais impessoal e o que recolhe informação menos qualificada. Os dados produzidos são armazenados numa base de dados, que serve de suporte à realização de campanhas de marketing específicas para a venda de serviço fixo. Porém, a realização dessas campanhas depende da capacidade disponível dos parceiros de telemarketing.

Sabe-se que a empresa trabalha com 3 parceiros, que praticam condições de contacto e preços diferentes no que toca a campanhas direcionadas para *leads.* Tipicamente, as campanhas de *leads* têm uma taxa de conversão muito baixa, pois trata-se de uma oferta muito específica e é necessário reunir um conjunto de condições para que ocorra a venda. Sendo o *telemarketing* um canal comissionado, isto é, os assistentes recebem comissões face ao número de vendas que realizam, esta torna-se uma situação pouco vantajosa para eles. Por esse motivo, cada um dos parceiros estabelece um limite máximo de contactos de *leads* a realizar por mês, assim como custos de contacto mais agressivos. Na tabela seguinte apresentam-se as características de cada um dos parceiros.

Tabela 6.1 - Número máximo de contactos e custo por contacto praticado por cada Parceiro de Telemarketing.

Parceiro de Telemarketing	Número Máximo de Contactos (contactos/mês)	Custo por Contacto (€/contacto)
Parceiro A	15 000	3€
Parceiro B	25 000	4€
Parceiro C	45 000	6€

Como se pode observar na tabela 6.1, o Parceiro A é o que pratica um custo por contacto menor, porém, também tem uma capacidade máxima de contactos mais reduzida. O parceiro C é aquele que disponibiliza mais recursos para campanhas de *leads*, praticando um custo unitário maior. Do ponto de vista de negócio, a necessidade de contactos por mês também é variável. Deste modo, fica ao critério da empresa selecionar o número de contactos a realizar por mês, face ao investimento alocado para aquele mês, ou face às metas que conversão que ambicionam.

Neste sentido, torna-se ainda mais essencial entregar aos parceiros contactos com maior qualidade e com maior probabilidade de conversão. No entanto, o processo de segmentação instituído era realizado manualmente, através da seleção de alguns atributos e segundo alguns critérios, como por exemplo a data de criação mais recente, o preenchimento de alguns campos como a data de final de contrato e a existência de cobertura de fibra na morada indicada. Como se verificou, o processo era altamente ineficiente e não havia muita confiança nos dados segmentados, o que se traduzia em taxas de sucesso muito baixas.

Assim, com esta investigação, o objetivo era construir um mecanismo que tornasse o processo de segmentação mais automático e inteligente, capaz de identificar e selecionar o conjunto de *leads* mais valioso e com maior propensão para converter, e ainda, com a flexibilidade de personalizar o número de contactos de acordo com as necessidades da empresa. O intuito é identificar regras, tendências e padrões de comportamento nos dados que sejam indicadores de possíveis oportunidades de conversão, para que, desse modo, se adote uma intervenção proactiva, no momento certo, reduzindo o número de contactos necessários. Idealmente, este modelo de avaliação e seleção de *leads* permitiria aumentar a eficiência e eficácia de futuras campanhas.

Ainda na perspetiva da empresa, esta investigação tem também a finalidade de fazer uma análise completa dos processos intermédios, de forma a identificar oportunidades de melhoria que possam trazer impacto na taxa de conversão final.

6.2.1. Análise e Identificação de Problemas

De modo a construir um conjunto de soluções integradas e alinhadas com as necessidades e com modelo de funcionamento do negócio, foi necessário realizar uma análise profunda dos processos e das suas dependências. Conforme especificado anteriormente, existem 3 fontes principais onde são produzidas e capturadas *leads*. As lojas e o telemarketing são canais assistidos, nos quais os assistentes são incentivados a criarem uma *leads* sempre que houver potencial interesse na aquisição de um serviço fixo. No canal digital, o *website*, a captura da *lead* ocorre através de um processo de verificação de cobertura, no qual o utilizador introduz os seus dados de contacto e morada com o intuito de verificar se é elegível para o serviço fixo. Nos casos em que existe cobertura, o utilizador acorda em receber uma chamada de seguimento. Através dessa chamada, efetuada pelo telemarketing, há oportunidade de fechar uma venda ou adicionar informação relevante à *lead.* Para o negócio, essa informação é fundamental para identificar a melhor altura para voltar a contactar, personalizar a oferta, ou simplesmente acompanhar o desfecho de cada interação anterior.

Quando uma *lead* é criada, independentemente do canal, é registada uma data de criação, a fonte, o nome do utilizador que a criou, e é definido um estado como "Nova". Existe também um campo automático que diz respeito à data de atualização, sendo que no momento inicial é igual à data de criação. Quando existe uma atualização da *lead*, toda os campos ficam editáveis exceto os mencionados anteriormente. O estado da *lead*, porém, tem de ser alterado manualmente, podendo registar os valores "Em Progresso", "Contactado" ou "Fechado".

Embora exista a possibilidade de editar e adicionar informação, a empresa reconhece que o foco sempre foi a captura massiva de *leads.* No entanto, nesta fase surge a consciência de que para muitas a qualificação é ainda prematura, não existindo sustentação para avançar com campanhas. Nesse sentido foram analisados todos os processos de captura, tratamento, contacto e as várias iniciativas a decorrer em paralelo. Foram também analisados os dados produzidos e o tipo de informação armazenada. Desta análise resultou a identificação de um conjunto de 11 problemas, que podem ser divididos nas várias fases sugeridas pelo modelo de gestão de *leads* apresentado no capítulo 3.

Captura

1- Reduzido número de *leads* produzidas pelos canais de vendas – na realidade nos canais de vendas, lojas e telemarketing, os assistentes são comissionados pelas vendas que realizam, não existindo nenhum incentivo ou motivação para criar *leads* ou atualizar a informação.

2- Elevado número de *leads* com dados inválidos, incorretos ou inconsistentes – segundo os dados observados, 90% das *leads* são capturadas através do *website*. Neste canal o comportamento do utilizador é mais volátil, sendo bastante propício o registo de números de telefone inválidos, como o típico 912345678. É também muito frequente o preenchimento de números aleatórios. Relativamente à morada, há por vezes inconsistências a nível da rua e do código postal.

3- Duplicação de *leads* no mesmo canal – é bastante frequente a existência de *leads* com exatamente a mesma informação, inclusive a data e hora de criação. Este comportamento pode ser resultante de erros ou duplos cliques.

Enriquecer

4- Duplicação de *leads* para o mesmo utilizador – este erro pode ocorrer de variadas formas, *online* quando o utilizador executa várias vezes o pedido de elegibilidade, ou nos canais assistidos devido à complexidade do processo de atualização. Como a procura e edição da *lead* carece de mais tempo, os assistentes optam por criar uma nova *lead* com a mesma informação.

5- 90% das *leads* são incompletas – quanto mais informação for coletada, maior é a qualificação da *lead.* Atualmente a esmagadora maioria das *leads* não satisfaz os requisitos mínimos de qualificação, que inclui o registo da data final de contrato, CED (do inglês *Contract End Date*).

6- Desatualização do campo "Elegibilidade" – este campo indica se, no momento de criação da *lead*, existia cobertura de fibra na morada indicada. No entanto, tem se verificado uma expansão considerável da área de cobertura, não existindo tal reflexão na base de dados. Esta falha origina perda de oportunidades e registo de conversões em *leads* marcadas como não elegíveis.

Nutrição

7- Falta de processo de acompanhamento para *leads* qualificadas – *leads* com registo de data de final de contrato são *leads* com elevado interesse e potencial. Esta informação permite contactar no momento certo, quando o prazo de fidelização está prestes a terminar. No entanto, não há procedimentos que visem o acompanhamento e nutrição destas *leads* ao longo do tempo, na tentativa de criar uma relação com o utilizador.

Rastrear

8- Atualização manual do estado – o campo que define o estado da *lead* é atualizado manualmente, originando que não exista um acompanhamento real da evolução. No conjunto de dados inicial, 94% das *leads* registavam um estado igual a "Nova".

Avaliar

9- Falta de métricas de desempenho – Não existe nenhum KPI (sigla do termo inglês *Key Performance Indicator)* instituído para medir o sucesso de cada campanha em tempo real.

10- Sem procedimentos de auditoria a avaliação da qualidade dos dados.

11- Falta de prioritização das *leads* com base no seu valor – não existe nenhum procedimento que indique o valor de cada *lead* ou indique a sua propensão para converter. Este processo é considerado essencial para rentabilizar e tornar mais eficiente o processo global.

Identificados os principais problemas a nível de processos e de solução técnica, segue-se a descrição do desenvolvimento dos modelos de limpeza e propensão, que visam responder aos problemas identificados nos pontos 10 e 11, respetivamente. Para os restantes pontos, a proposta de solução é apresentada no capítulo 6.5.1, com as respetivas recomendações de melhoria da componente prática da investigação.

6.3. Pré-processamento dos dados – Modelo de Limpeza

Como resultado de uma primeira análise dos dados e dos processos de captura e atualização dos dados, sentiu-se a necessidade de recorrer a um processamento prévio dos dados. De facto, considerou-se que a qualidade do conjunto inicial de dados não era satisfatória para a construção do modelo de propensão, uma vez que seriam expostos muitos erros, ruídos e padrões incorretos.

Como tal, conceptualizou-se um algoritmo que resultou da sequência lógica de um conjunto de regras. Essas regras foram inferidas junto das áreas de negócio, e também dos problemas identificados anteriormente. Obteve-se, então, num conjunto de 18 regras que, aplicadas sequencialmente, avaliam com base em critérios de negócio quais os registos que devem ser excluídos do conjunto de dados. Idealmente, a aplicação deste modelo deverá sempre preceder o modelo de propensão, garantido que a aprendizagem e a segmentação são realizadas sobre uma base de dados qualificada.

De seguida, são apresentados os critérios aplicados, agrupados por fatores ou categorias de decisão. Note-se que por questões de confidencialidade dos dados não puderam ser partilhadas partes do código. Apresenta-se, no entanto, o respetivo impacto de cada conjunto de critérios na limpeza dos dados, através do número de registos excluídos em cada passo. O conjunto inicial era composto por 1 277 327 registos e 18 variáveis.

Tabela 6.2 - Conjunto de critérios aplicados na construção do modelo de limpeza.

Ordem	Descrição do Critério de Exclusão	*Leads* Excluídas
1º	*Leads* duplicadas, com exatamente a mesma informação em todas as variáveis.	222 122 **(17%)**
2º	Normalização das moradas. *Leads* duplicadas para a mesma morada.	2 965 **(0,2%)**
3º	Para *leads* com a mesma morada, excluir aquelas menos qualificadas, com base na identificação do decisor.	140 615 **(11%)**
4º e 5º	*Leads* pouco qualificadas, cujo número de telefone e a morada aparecem em mais de 10 combinações diferentes destas duas variáveis.	205 022 **(16%)**
6º a 10º	Para a mesma pessoa, excluir as *leads* menos qualificadas com base na CED, canal de criação, data de criação mais antiga, data de atualização mais antiga.	147 686 **(12%)**
11º a 13º	*Leads* com diferentes CED, desatualizadas e repetidas, valorizando os mais recentes e mais frequentes.	2 393 **(0,2%)**
14º a 18º	*Leads* com a mesma informação, excluir as mais antigas, as mais desatualizadas ou menos qualificadas.	10 107 **(0,8%)**

Deste modo, o encadeamento dos 18 critérios de exclusão permitiu remover 730 910 *leads*. Estas *leads* foram excluídas do conjunto de dados para análise por serem consideradas irrelevante ou pouco qualificadas. Para a maior parte das *leads*, referentes a uma mesma pessoa ou residência, existia sempre outra *leads* com mais informação, ou criada por um canal comercial, ou mais recente, ou mais atualizada, tornando possível reduzir o volume de dados.

Como foi abordado no capítulo 4, a variabilidade e heterogeneidade dos dados é brutal de empresa para empresa, e até entre vários departamentos dentro de uma mesma organização. E como tal, não há um modelo de limpeza ótimo e que se aplique a todos os casos. Existe sim, no entanto, um conjunto de técnicas e procedimentos que facilitam a identificação de erros e ruídos que afetam a qualidade dos dados. O modelo de limpeza deve ser desenhado à medida das necessidades e das regras de negócio de cada empresa, tendo por base o estudo estatístico e analítico dos fatores de erro. Assim, foi possível reduzir em 57% o conjunto de dados inicial.

6.4. Modelo de Propensão de *Leads*

O presente capítulo foi dedicado à construção do modelo de propensão de *leads* para adesão de serviço fixo, aplicando as diversas técnicas descritas no capítulo 4. Como descrito anteriormente, a metodologia adotada resulta na combinação das duas estratégias de Descoberta de Conhecimento em Base de Dados. Nas secções anteriores foi exercida a metodologia CRISP-DM, através da caracterização dos negócio e especificação dos requisitos técnicos e funcionais. De seguida, aplicou-se a metodologia SEMMA para a construção do modelo, seguindo as etapas recomendadas: Exploração dos dados, Processamento dos dados, Modelação e Comparação e Avaliação dos resultados.

A metodologia SEMMA foi integralmente desenvolvida recorrendo à ferramenta *SAS Enterprise Miner*. O *software* permite agilizar o processo de *data mining* ao criar modelos preditivos altamente precisos com base na análise de grandes quantidades de dados. A construção modular do modelo e a simples parametrização de variáveis tornou o desenvolvimento bastante intuitivo e flexível. O único processamento que foi necessário realizar fora da ferramenta, foi a preparação dos conjuntos de dados.

Inicialmente, como resultado do modelo de limpeza, resultaram 546 417 registos. Destes, apenas 103 881 tinham uma classificação conhecida, isto é, existia registo de sucesso ou insucesso na conversão. Desse modo, produziu-se um conjunto de dados com os 103 881 registos, que seriam utilizados para o conjunto de treino, e os restantes 442 536 registos foram posteriormente utilizados para aplicar as técnicas de classificação. O objetivo é induzir regras e padrões escondidos nos dados de treino que permitam identificar e selecionar destes 442 536 registos, as *leads* com maior propensão para aderir ao serviço fixo, tornando o processo de segmentação menos arbitrário e facilitando a tomada de decisão.

6.4.1. Exploração dos Dados

Posteriormente à preparação das amostras representativas dos dados, seguiu-se a fase de exploração dos dados. Conforme descrito na secção 4.3.2, esta fase tem como objetivo a concretização de uma análise qualitativa e quantitativa dos dados, com o intuito de visualizar e antecipar alguns comportamentos, tendências e anomalias, facilitando a compreensão das características em estudo. Para tal, recorreu-se à importação dos dados para o *software SAS Enterprise Miner* e aos seus componentes de exploração (*StatExplore, Graph Explore, Multiplot e Variable Clustering*) como se pode observar na figura 6.1.

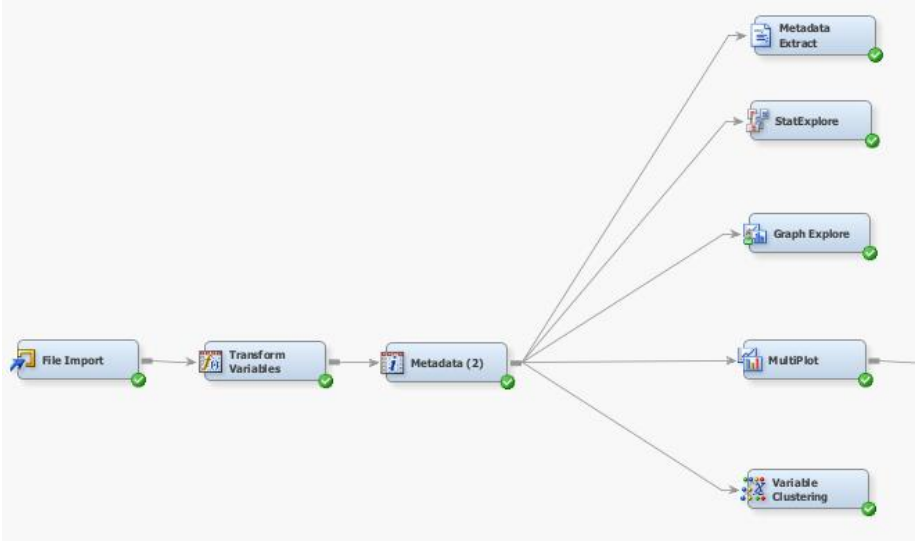

Figura 6.1 - *SAS Enterprise Miner* - Componentes de Exploração dos Dados.

Através dos módulos aplicados pode apurar-se que o conjunto de dados inicial era composto por 103 881 linhas e um total de 17 variáveis, o que constitui o essencial da informação recolhida na criação de uma *lead.* Na tabela 6.3 apresenta a lista de variáveis, com a respetiva descrição, papel e tipo de variável.

Tabela 6.3 - Lista de variáveis consideradas e respetiva descrição, papel e tipo.

Variável	Descrição	Função	Tipo
Ced_Dt	Indicador da existência de data de final de contrato na data de criação da lead.	*Input*	Nominal
Conversion	Variável dependente, indica se houve conversão final da lead.	*Target*	Binária
Count_Phone	Número de vezes que o número de telefone indicado na lead foi utilizado com moradas diferentes.	*Input*	Intervalar
Count_Ua	Número de vezes que a morada indicada na lead foi utilizada com números de telefone diferentes.	*Input*	Intervalar
Decision_Maker	Indica se o número de telefone utilizado na criação da lead corresponde ao decisor na morada.	*Input*	Nominal
Dif_Ced_Update	Diferença, em meses, entre a data de última atualização e data de final de contrato da lead.	*Input*	Intervalar
Dif_Creation_Update	Diferença, em dias, entre a data de criação e data de última atualização da lead.	*Input*	Intervalar
Distrito	Distrito associado ao código postal indicado na lead.	*Input*	Nominal
Dw_Lead_Creation_Dt	Data de criação da lead, no formato AAAAMMDD.	*Input*	Intervalar
E_Lead_Id	Identificador único da lead.	*ID*	Intervalar
Elegivel	Indicador de elegibilidade no momento de criação da lead.	*Input*	Nominal
Estado_Lead	Estado de atualização da lead.	*Input*	Nominal
Lead_Level	Identificação do nível da lead.	*Input*	Nominal
Lead_Source	Canal de criação da lead	*Input*	Nominal
Lead_Update	Indicador de atualização da lead posteriormente à sua criação.	*Input*	Nominal
Phone_Type	Tipo de número de telefone fornecido na criação da lead.	*Input*	Nominal
Postal_Code	Código postal indicado na lead.	*Input*	Nominal

Com o intuito de proteger a informação pessoal e privacidade dos clientes, algumas variáveis sofreram alterações de forma a mascarar os dados e em alguns casos, facilitar a deteção de padrões. Como descrito na tabela anterior, no estudo da variável dependente foram consideradas 10 variáveis nominais e 5 variáveis intervalares. A variável dependente *Conversion* indica se, posteriormente à criação da *lead,* foi registada uma nova ativação de serviço fixo na morada ou número fiscal indicado na *lead.* Do conjunto das 103 881 leads disponibilizadas, 58 924 converteram com sucesso (valor igual a 1), o que representa cerca de 57% do conjunto de dados, e as restantes 44 957 rejeitaram a oferta (valor igual a 0), dando como fechadas os outros 43% dos registo.

Analisando a distribuição das variáveis nominais, apresentadas na figura 6.2, pode verificar-se quais os valores mais frequentes e qual o número de níveis para cada variável. A data de final de contrato (CED_DT), por exemplo, apresenta apenas dois níveis possíveis, Y e N, sendo que nível mais frequente é N, que ocorre em 80.98% dos registos do conjunto de treino. O estado da *lead* tem 4 níveis, admitindo os valores Nova, Em Progresso, Contactado ou Fechado, sendo que a maior incidência ocorre no estado Em Progresso (51,47%), como se pode verificar na figura 6.7. Variáveis como o código postal e o distrito são as que apresentam maior variabilidade, visto que são variáveis de localização. Interessante observar que 29,83% das *leads* foram criadas no distrito de Lisboa e 27,65% no Porto. O canal com maior potencial para captura de *leads* é o site *online,* como seria expectável.

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	CED_DT	INPUT	2	0	N	80.98	Y	19.02
TRAIN	DECISION_MAKER	INPUT	2	0	N	87.75	Y	12.25
TRAIN	DISTRITO	INPUT	20	0	Lisboa	29.83	Porto	27.65
TRAIN	ELEGIVEL	INPUT	2	0	Y	96.81	N	3.19
TRAIN	ESTADO_LEAD	INPUT	4	0	EM PROGRESSO	51.47	NOVA	48.16
TRAIN	LEAD_LEVEL	INPUT	3	0	Independent Lead	56.10	Client	43.46
TRAIN	LEAD_SOURCE	INPUT	8	0		70.19	VFE Telemarketing	19.98
TRAIN	LEAD_UPDATE	INPUT	2	0	Y	51.81	N	48.19
TRAIN	PHONE_TYPE	INPUT	2	0	MOVEL	85.68	FIXO	14.32
TRAIN	POSTAL_CODE	INPUT	432	0	4400	1.35	2560	1.27
TRAIN	CONVERSION	TARGET	2	0	1	56.72	0	43.28

Figura 6.2 - *SAS Enterprise Miner* - Estatísticas básicas das variáveis nominais.

Sabendo que a maioria das *leads* são criadas no contexto de verificação de cobertura no *website*, verifica-se que 96,81% das *leads* eram elegíveis para instalação do serviço fixo. Uma vez que houve o tratamento prévio dos dados, todos os registos deste conjunto não apresentam *missing values*, não sendo necessário recorrer aos métodos descritos na secção 4.4.1.

Analisando as estatísticas básicas das 5 variáveis intervalares é possível ter uma perceção das respetivas distribuições, embora não seja expectável que sigam nenhuma distribuição probabilística conhecida. Na figura 6.3, para além das estatísticas mais conhecidas apresentam-se também os coeficientes de Skewness e Kurtosis. O valor do coeficiente de Skewness indica a assimetria de uma variável, sendo que valores negativos traduzem uma cauda para a esquerda, enquanto que valores positivos traduzem uma cauda para a direita. Valores de Skewness próximos de 0 indicam uma simetria em relação ao valor médio. O coeficiente de Kurtosis refere-se também a uma característica gráfica da distribuição, o achatamento da curva.

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
COUNT_PHONE	INPUT	1.640656	1.336728	103881	0	1	1	9	2.966186	9.851256
COUNT_UA	INPUT	1.307121	0.775191	103881	0	1	1	12	3.943022	21.77913
DIF_CED_UPDATE	INPUT	-80928	39359.76	103881	0	-100000	-100000	4100	1.579373	0.494868
DIF_CREATION_UPDATE	INPUT	2261.71	7059.45	103881	0	0	300	57600	4.161441	17.56261
DW_LEAD_CREATION_DT_ID	INPUT	20160634	5120.224	103881	0	20150719	20160620	20170312	-0.01322	27.48863

Figura 6.3 - *SAS Enterprise Miner* – Estatísticas básicas das variáveis intervalares.

O valor referência destes coeficientes aplica-se à distribuição Normal, com fator de Kurtosis igual a 3 e Skewness igual a 0. Como se pode constatar, nenhuma variável aparenta ser normalmente distribuída.

Segundo a figura acima representada, pode verificar-se que as variáveis Count_Phone e Count_Ua, que representam o número de vezes que um número de telefone foi pesquisado com moradas diferentes e vice-versa, têm distribuições semelhantes. Na figura 6.4 apresentam-se os histogramas referentes às duas variáveis, onde se pode confirmar este comportamento, embora o Count_Ua tenha uma amplitude de valores ligeiramente maior.

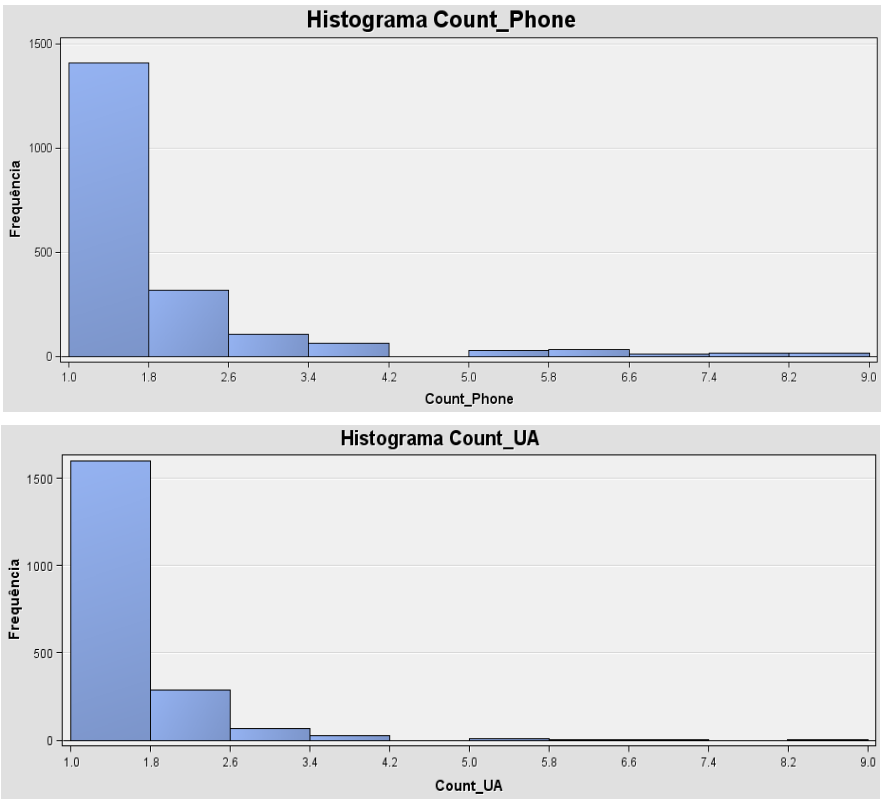

Figura 6.4 - *SAS Enterprise Miner* - Histogramas das variáveis Count_Phone e Count_Ua.

O número de meses entre a data de final de contrato e a última atualização (Dif_CED_Update) varia entre -1000 e 41, sendo que -100 000 é indicador que a *lead* não tinha data de final de contrato registada. A variável Dif_Creation_Update, que representa o número de dias entre a data de criação e a data de atualização, varia entre 0 e 57 600, sendo 0 os casos em que a *lead* não sofreu qualquer atualização desde a sua criação. Quanto à data de criação, pode verificar-se que a mais antiga data a 19 de Julho de 2015 e a mais recente 12 de Março de 2017.

Recorrendo ao módulo *StatExplore* do *SAS Miner*, produziu-se um gráfico que demonstra, por ordem decrescente, o valor que cada variável tem na determinação da variável dependente. Através da figura 6.5 pode constatar-se que a diferença entre data de criação e de atualização é, de facto, a variável com maior importância e que melhor explica a conversão de uma *lead*, com um valor de 0.19, enquanto que o tipo de telefone é a que menos contribui para explicar a variável dependente.

De modo a investigar o comportamento da conversão de *leads* face a cada uma das variáveis, recorreu-se às várias técnicas representadas na figura 6.1, seguindo uma lógica de prioritização das variáveis com maior valor.

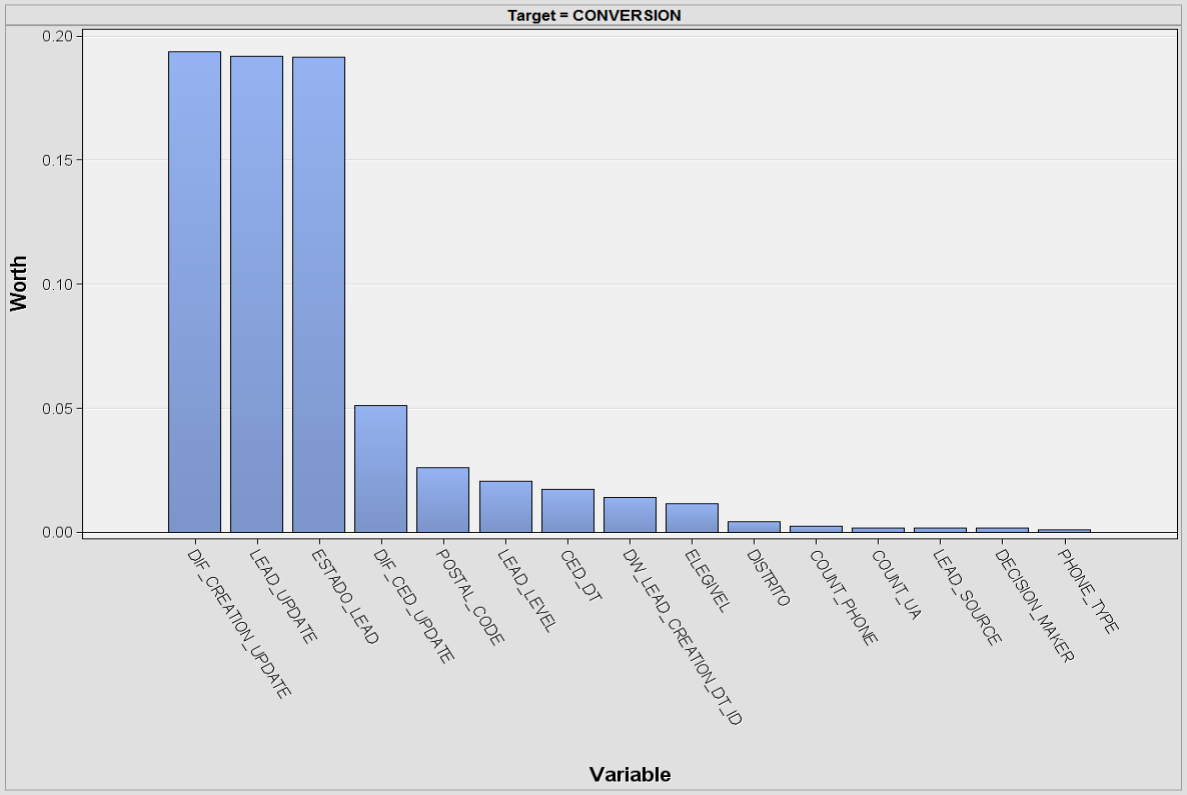

Figura 6.5 - *SAS Enterprise Miner* - Representação do valor de cada variável face à variável dependente.

Analisando a variável Lead_Update na figura 6.6, verifica-se que que a amostra é bastante equilibrada no que diz respeito a esta variável, uma vez que cerca de 50% das *leads* sofreram atualização. No entanto, os dados demonstram um comportamento inesperado: apenas 27,3% das *leads* que foram atualizadas converteram com sucesso, enquanto que 88,4% das que não sofreram um segundo contacto obtiveram conversão. Esta ocorrência deve-se ao facto de a atualização depender da inserção manual dos dados, não refletindo o verdadeiro comportamento. Quando num segundo contacto a *lead* demonstra interesse em converter, não é dada prioridade à atualização dos dados da *lead*, produzindo estes comportamentos inconsistentes.

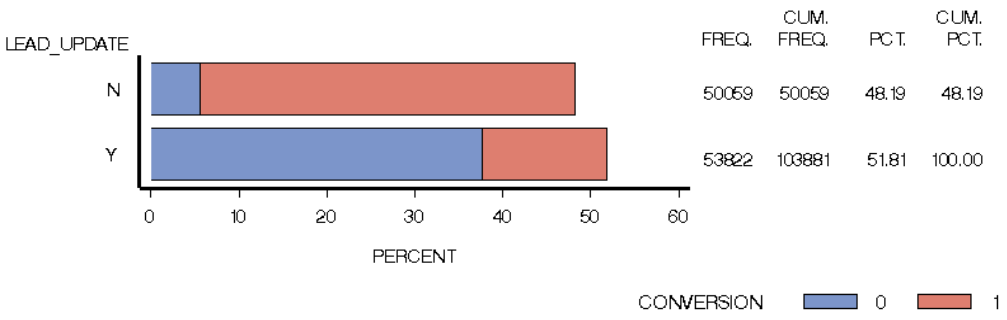

Figura 6.6 - *SAS Enterprise Miner* - Distribuição da variável independente, Lead_Update, face à variável dependente, Conversão.

A terceira variável com maior relevância é o estado da *lead*. Como se pode observar na figura 6.7, há um grande desequilíbrio na amostra, pois 99,6% correspondem a *leads* no estado Nova ou Em Progresso, enquanto que os restantes 0,4% são distribuídos entre Contactado ou Fechado. Havendo uma elevada taxa de conversão, era esperado um maior número de leads nos dois últimos estados. Mais uma vez, os dados representam inconsistências devido à dependência de atualização manual dos estados. No capítulo 6.5.1 apresentam-se propostas de solução para evitar estas situações e, como consequência, melhorar a qualidade dos dados.

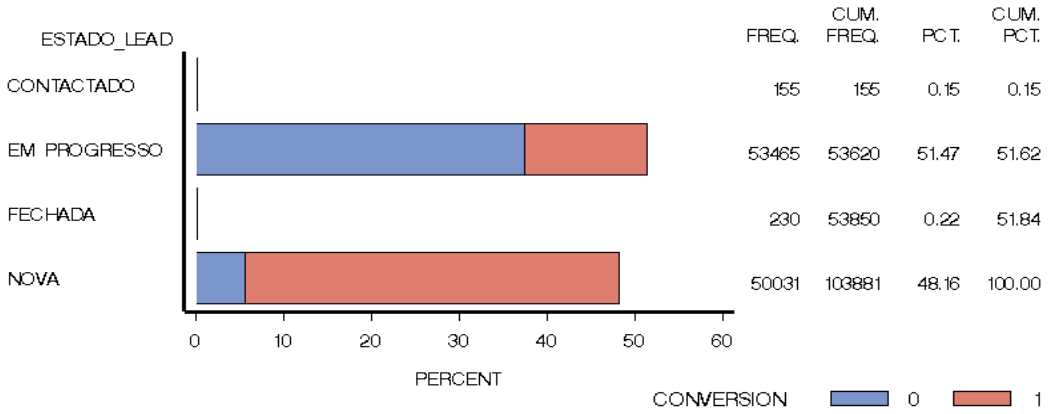

Figura 6.7 - *SAS Enterprise Miner* - Distribuição da variável independente, Estado_Lead, face à variável dependente, Conversão.

Representada na figura 6.8, a variável Lead_Level identifica se a *lead* foi criada por um cliente da operadora ou por um não cliente. No caso de não clientes existe um nível específico, o *Prospect*, que identifica indivíduos que estabeleceram contacto através de um canal assistido, mostrando um interesse particular, que foi registado pelo próprio assistente. Conforme é apresentado na figura seguinte, 56% das *leads* foram criadas por não clientes, o que traduz uma grande oportunidade de negócio para a operadora, no sentido em que pode aumentar a sua quota de mercado. Por outro lado, os 43% de clientes que registaram o seu interesse ao verificar a cobertura da sua residência representam uma oportunidade de *upselling*, ou seja, aumentar o valor de receita do cliente, apresentando, por exemplo, planos convergentes. Nesse sentido, verifica-se uma taxa de sucesso de 68,3% na estratégia de *upselling*.

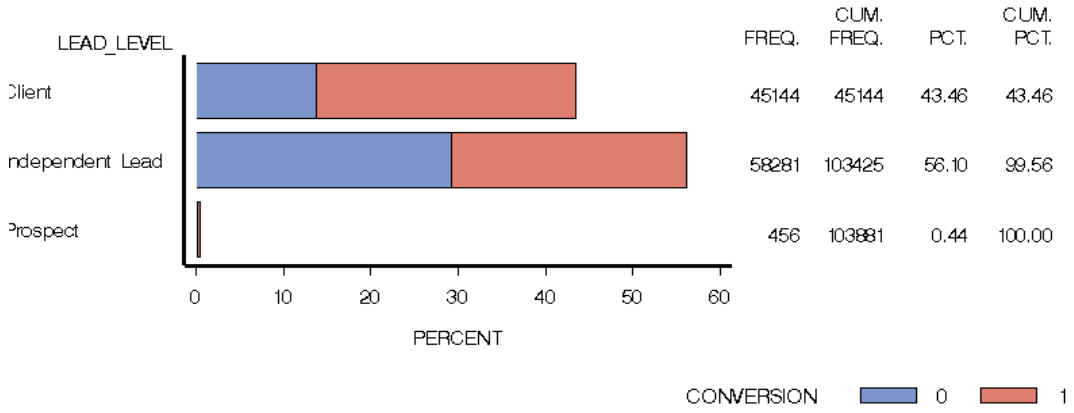

Figura 6.8 - *SAS Enterprise Miner* - Distribuição da variável independente, Lead_Level, face à variável dependente, Conversão.

A data de final de contrato (CED) é um indicador valioso na recolha de *leads*, contudo, uma informação difícil de obter e que carece de introdução manual no sistema informático. Como se pode observar na figura 6.9, cerca de 81% das *leads* do conjunto de dados não têm CED. Destas, porém, 61% converteram com sucesso, o que indica que, apesar de ser uma informação relevante, não é determinante para a conversão da *lead*.

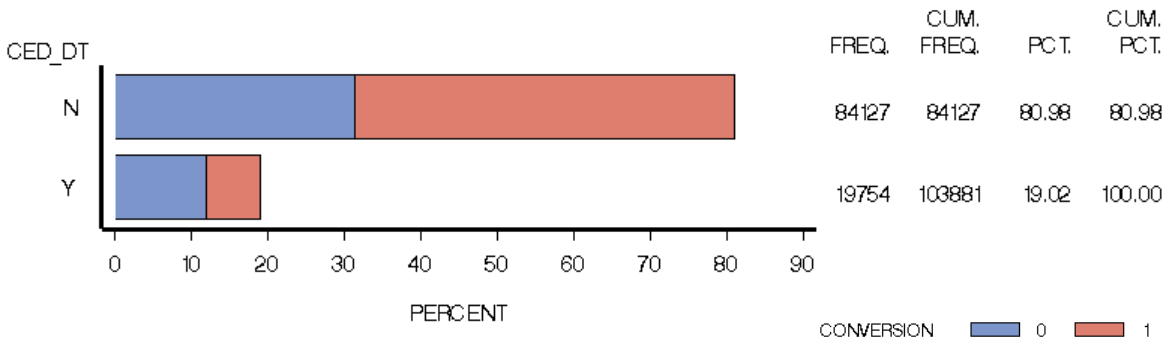

Figura 6.9 - *SAS Enterprise Miner* - Distribuição da variável independente, CED_DT, face à variável dependente, Conversão.

Porém, a elegibilidade é determinística na conclusão do contacto da *lead*, uma vez que não pode haver conversão com sucesso numa morada sem cobertura. Do total de registos, 97% apresentam cobertura na morada indicada, como se pode concluir através da figura 6.10.

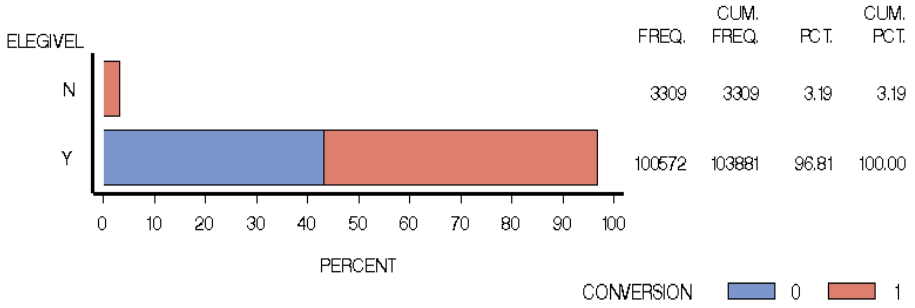

Figura 6.10 - *SAS Enterprise Miner* - Distribuição da variável independente, Elegivel, face à variável dependente, Conversão.

Através das pesquisas de elegibilidade, é possível determinar quais as zonas geográficas mais procuradas. Na figura 6.11 verifica-se uma maior concentração nas regiões de Lisboa e Porto, mas também com grande presença nos distritos de Aveiro e Braga.

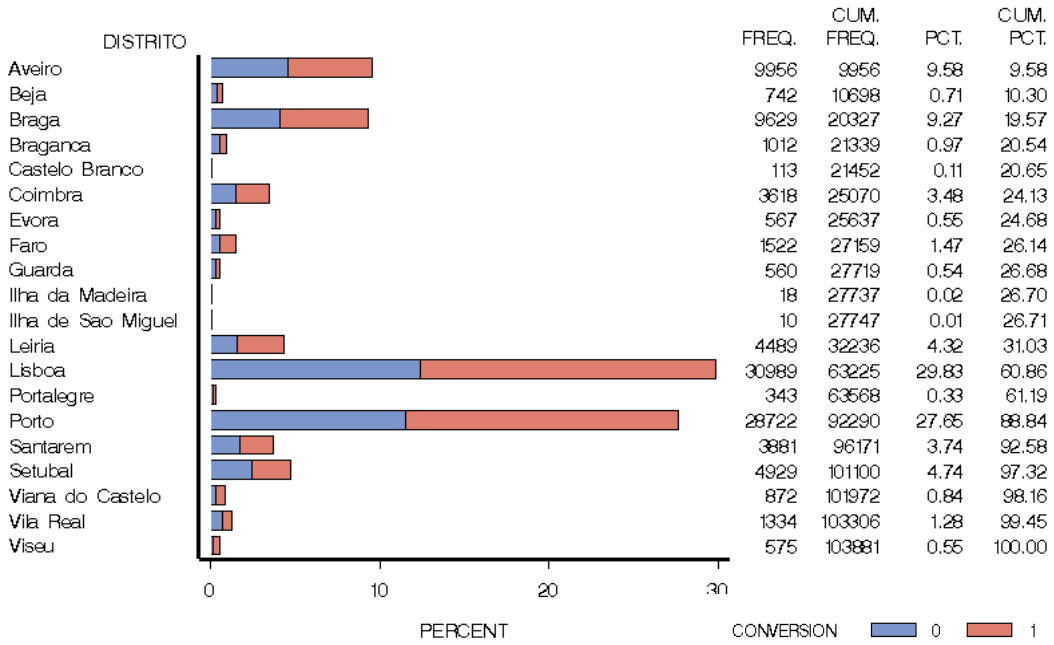

Figura 6.11 - *SAS Enterprise Miner* - Distribuição da variável independente, Distrito, face à variável dependente, Conversão.

Nas ilhas da Madeira e São Miguel o número de *leads* é residual, no entanto, a taxa de conversão é 100%, o que leva a identificar um caso em que as *leads* são criadas na eminência de uma conversão e não numa lógica de angariação de uma base de contactos com potencial interesse. Das restantes

localidades, Viseu, Leiria e Faro são os distritos que apresentam maior taxa de conversão, na ordem dos 68%, 64% e 63%, respetivamente. Castelo Branco, por sua vez, tem a menor taxa de conversão de *leads* (20,35%).

As variáveis Count_Phone e Count_UA são indicadores do nível de confiança de uma *lead* no que diz respeito à combinação número de telefone e morada. À partida será mais difícil oferecer a campanha certa a um indivíduo que utilizou o número de telefone para pesquisar diferentes moradas. Dá-se o exemplo em que um utilizador do *website* decide verificar a cobertura na sua morada, para a qual ainda não está disponível o serviço de fibra. No entanto, o utilizador continua a sua pesquisa, procurando pela casa de um familiar ou outras casas na vizinhança. Neste exemplo, torna-se complexo a atribuição da morada correta ao utilizador. O mesmo acontece quando uma morada é pesquisada por diversos contactos, dificultando a identificação do decisor real. Na figura 6.12 são apresentados os gráficos de barras para ambas as variáveis, onde se verifica que a grande maioria dos registos são únicos, ou seja, têm uma combinação de 1 para 1. Este facto deriva como resultado do modelo de limpeza aplicado anteriormente.

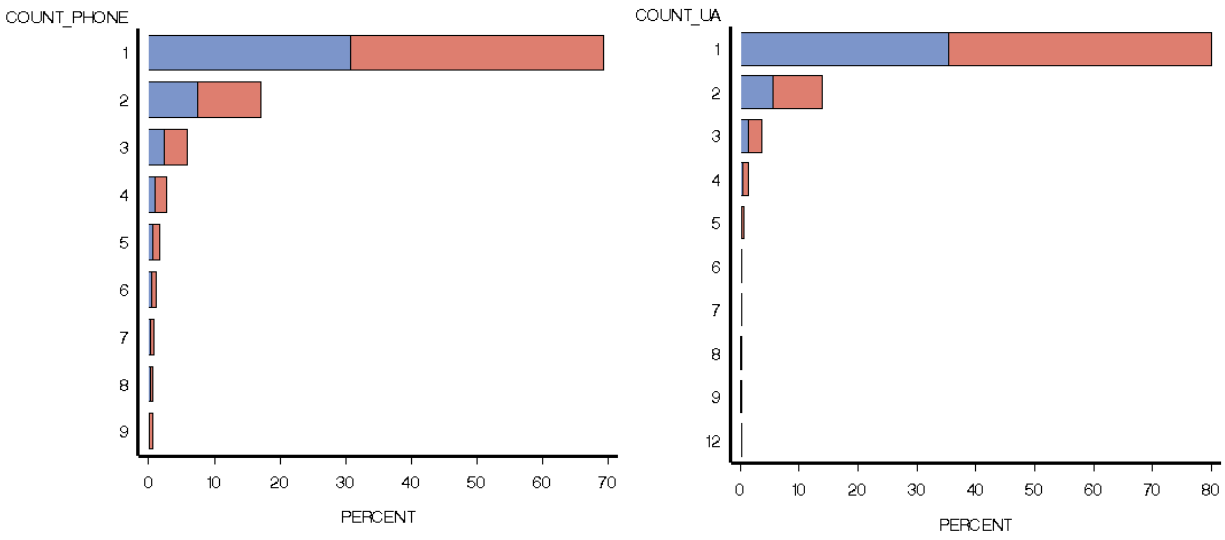

Figura 6.12 - *SAS Enterprise Miner* - Distribuição das variáveis independentes, Count_Phone e Count_Ua, face à variável dependente, Conversão.

Examinando os diversos canais de recolha de *leads* pode apurar-se que o grande volume de criação e conversão é atribuído ao meio digital. O canal de *telemarketing*, apesar de significativo, trabalha essencialmente em *leads* originalmente criadas no *website*, sendo um potenciador nas conversões deste canal. Observe-se a figura 6.13.

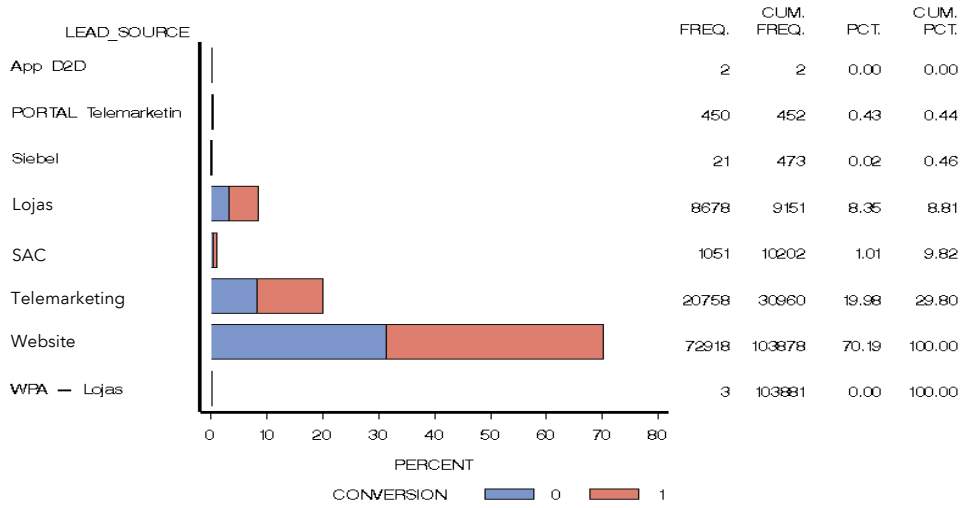

Figura 6.13 - *SAS Enterprise Miner* - Distribuição da variável independente, Lead_Source, face à variável dependente, Conversão.

Os dados de decisor são ainda prematuros, sendo complexo na maior parte dos casos, identificar o decisor para a morada mencionada. Embora não pareça significativo, a taxa de conversões nos decisores identificados é 64%, enquanto que para não decisores é 56%.

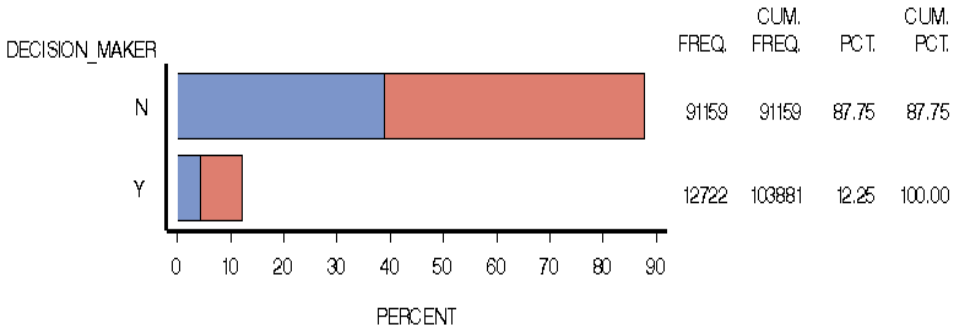

Figura 6.14 - *SAS Enterprise Miner* - Distribuição da variável independente, Lead_Source, face à variável dependente, Conversão.

Por fim, pode verificar-se que 85% das *leads* são criadas com um número de telefone móvel, o que demonstra uma preferência pelo contacto direto para o telefone. O número fixo, apensar de em menor escala, apresenta 50% de sucesso na conversão.

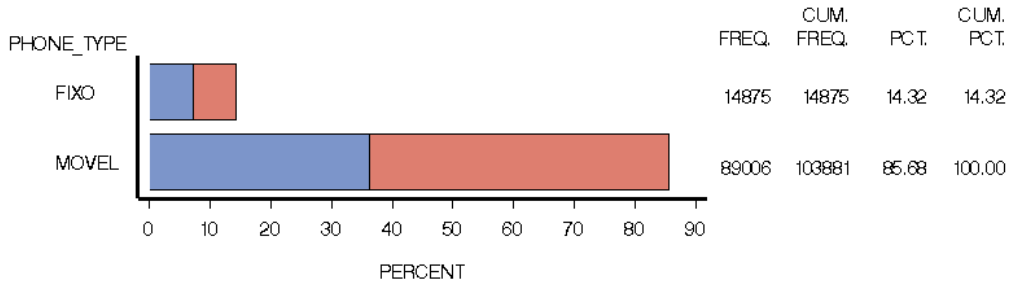

Figura 6.15 - *SAS Enterprise Miner* - Distribuição da variável independente, Phone_Type, face à variável dependente, Conversão.

6.4.2. Processamento dos Dados

No capítulo 4.4 refletiu-se sobre a importância do pré-processamento dos dados no processo de *data mining*, dado o enorme impacto que a qualidade dos dados tem nos resultados produzidos. No que diz respeito à qualidade de um modelo preditivo, esta depende em grande parte da qualidade do conjunto de dados utilizados na fase de aprendizagem.

Na secção anterior, onde foi realizada uma exploração qualitativa e quantitativa das variáveis consideradas, pôde constatar-se que os dados recolhidos no processo de criação e tratamento de *leads* são muitas vezes incompletos, inconsistentes e dependem do preenchimento ou atualização manual. Estes fatores prejudicam a qualidade dos dados, pois potenciam a introdução de erros na base de dados e refletem comportamentos que não correspondem à realidade. Porém, há várias técnicas de tratamento de dados que, aplicadas sobre os dados de treino, permitem tornar o conjunto de dados mais homogéneo, consistente e confiável.

Nas secções seguintes apresentam-se as técnicas de exclusão de *outliers*, transformação de variáveis, partição dos dados e seleção de variáveis aplicadas ao conjunto de dados, com o objetivo tornar a modelação mais eficaz e eficiente.

6.4.2.1. Limpeza dos Dados

Como mencionado no capítulo 4.4.1, muito do ruído dos dados é removido através da identificação de valores extremos, valores raros, erros e valores em falta. Como resultado da exploração gráfica e tabelar das variáveis nominais e intervalares foram detetados registos que, por erro ou por representarem casos pouco comuns, foram excluídos da amostra de treino. As regras de exclusão aplicadas foram as seguintes:

- Lead_Level = "*Prospect"*;
- Count_UA > 9;

Identificou-se que a amostra referente à primeira condição era residual, acrescentado variabilidade e complexidade desnecessária ao conjunto. Relativamente a *leads* cuja morada corresponde a mais do que 9 números de telefone diferentes, admitiu-se um nível de confiança demasiado reduzido na utilização destas *leads.* Na perspetiva de negócio, este são os casos em que se torna difícil dirigir uma oferta ao decisor correto, prejudicando a eficiência e resultados das campanhas.

Foram filtrados um total de 459 registos, o que representa 0,44% do conjunto de dados, não correndo o risco de eliminar um número significativo de *leads* pois é inferior a 3%.

No que diz respeito à existência de *missing values*, pôde apurar-se que todas as variáveis estavam completas. Contudo, este fenómeno resultou do tratamento prévio dos dados até à fase de importação para o *SAS Miner,* onde os campos vazios foram preenchidos com valores pré-definido. Assim, não foi necessário utilizar nenhum método de preenchimento ou exclusão, mas revelou-se necessário analisar o impacto destes casos na qualidade dos dados. As respetivas ações de limpeza foram aplicadas posteriormente à fase de transformação das variáveis.

6.4.2.2. Transformação de Variáveis

Na construção de um modelo preditivo, a qualidade do resultado produzido é condicionada pelas variáveis utilizadas, sendo da maior importância a correta seleção das mesmas. As variáveis devem ser representativas do perfil de cliente, mas devem também estar alinhadas com o problema em estudo e com as necessidades de negócio.

Recorrendo ao módulo de transformação de variáveis do *SAS Miner* foi possível criar novas variáveis, com o objetivo de reduzir variabilidade, tornar mais evidente certos padrões e relacionar variáveis

através de transformações matemáticas, o que facilita também o processo de redução de dimensionalidade. Foram então criadas 4 variáveis, que se apresentam resumidas na tabela 6.4.

Tabela 6.4 - Lista de novas variáveis introduzidas no modelo.

Variável	Descrição	Tipo
Days_Creation_Extract	Diferença, em dias, entre a data de criação da *lead* e a data de extração do conjunto de dados. Representa a longevidade da *lead no momento da extração.*	Intervalar
Update_Status	Estado de atualização da *lead.* Representa, face ao intervalo de meses entre a CED e a última atualização, o estado de contacto da *lead.*	Nominal
Sum_Counts	Soma das variáveis Count_Phone e Count_UA. Representa o nível de confiança na *lead*, sendo que quanto menor o valor, maior a confiança na qualidade dos dados.	Intervalar
Multi_Counts	Produto das variáveis Count_Phone e Count_UA. Representa o nível de confiança na *lead*, sendo que quanto menor o valor, maior a confiança na qualidade dos dados.	Intervalar

Como exibido na descrição da tabela anterior, a variável Days_Creation_Extract resulta da simplificação da data de criação da *lead*, comparando-a com a data fixa de extração dos dados. Esta transformação facilita a interpretação dos dados e reduz o número de variáveis necessárias. A variável Update_Status foi criada com objetivo de reduzir o número de níveis da variável DIF_CED_UPDATE, e pode tomar um dos cinco níveis:

- ***Enhance*** – quando não há informação suficiente sobre a *lead*, nomeadamente a data de final de contrato (CED).
- ***Too Late* –** quando a atualização da *lead* ocorre posteriormente à data de final de contrato.
- ***Late* –** quando a atualização da *lead* ocorre na data de final de contrato.
- ***Nurturing* –** quando a atualização da *lead* ocorre durante os últimos 3 meses de contrato.
- ***Contact* –** quando a última atualização da *lead* ocorre a mais de 3 meses do final de contrato.

Ao transformar a variável data, que é intervalar, numa variável categórica, esta transformação reduz a variabilidade e contribui para a rápida interpretação dos dados. Por fim, o Sum_Counts e o Multi_Counts representam agregações de variáveis existentes, com o objetivo de explorar relações entre variáveis e reduzir o número total de atributos em análise.

De seguida, procedeu-se à validação da coerência dos dados criados. Impôs-se também a fase de seleção manual das variáveis que prosseguiriam para a fase de modelação. Conforme identificado, cerca de 80% das *leads* foram criadas sem a data de final de contrato (CED), mesmo em casos onde se finalizou uma conversão. Apesar de fornecer uma informação valiosa do ponto de vista de negócio, esta é uma variável que levanta questões face à sua relevância e impacto na qualidade dos dados, uma vez que deixa de ser determinante para a conversão e introduz padrões incorretos nos dados. Nessa perspetiva, optou-se por remover do conjunto de dados as variáveis relacionadas com esta informação: CED_DT e DIF_CED_UPDATE.

O atributo DW_LEAD_CREATION_DT deu origem à variável transformada, pelo que foi removida de modo a evitar a introdução de redundância no modelo. Face ao número residual de *leads* não elegíveis, removeu-se também a variável ELEGIVEL. A elegibilidade, como foi referido, é um fator determinante na conversão de uma *lead.* No entanto, há alguns registos classificados como conversão com sucesso mesmo em *leads* sem elegibilidade, o que levanta questões relativamente à precisão deste mecanismo. Pode também dar-se o caso de a *lead* não ser elegível no momento de criação e, posteriormente, a cobertura da rede ser alargada, não refletindo essa informação na base de dados. Assim, considerou-se que a variável ELEGIVEL não iria acrescentar informação relevante e qualificada ao modelo, devendo a validação de cobertura ser realizada após o processo de segmentação de cada campanha, e imediatamente antes do contacto.

Por fim, com o intuito de validar as alterações efetuadas e garantir a coerência e qualidade dos dados, fez-se um nova análise gráfica e estatística dos atributos. Nas figuras seguintes são apresentadas as alterações substanciais. A figura 6.16 representa a distribuição da nova variável Update_Status face à variável dependente. A figura 6.17 representa a nova escala de valor de cada atributo.

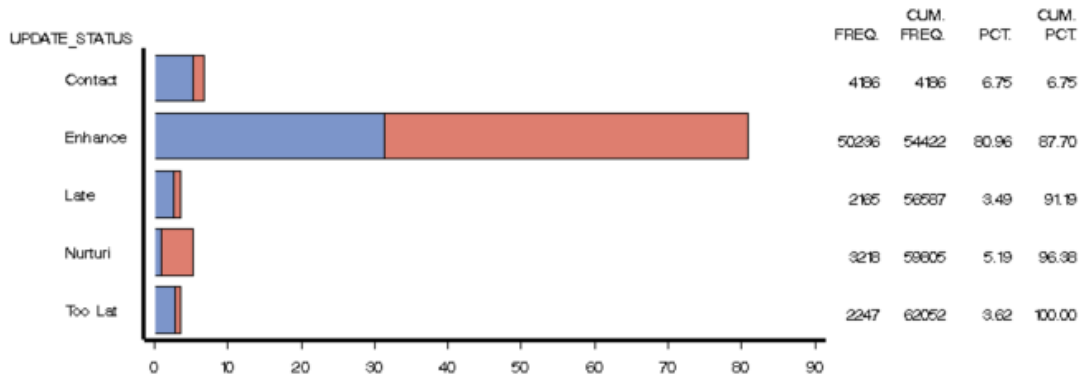

Figura 6.16 - *SAS Enterprise Miner* - Distribuição da variável independente, Update_Status, face à variável dependente, Conversão.

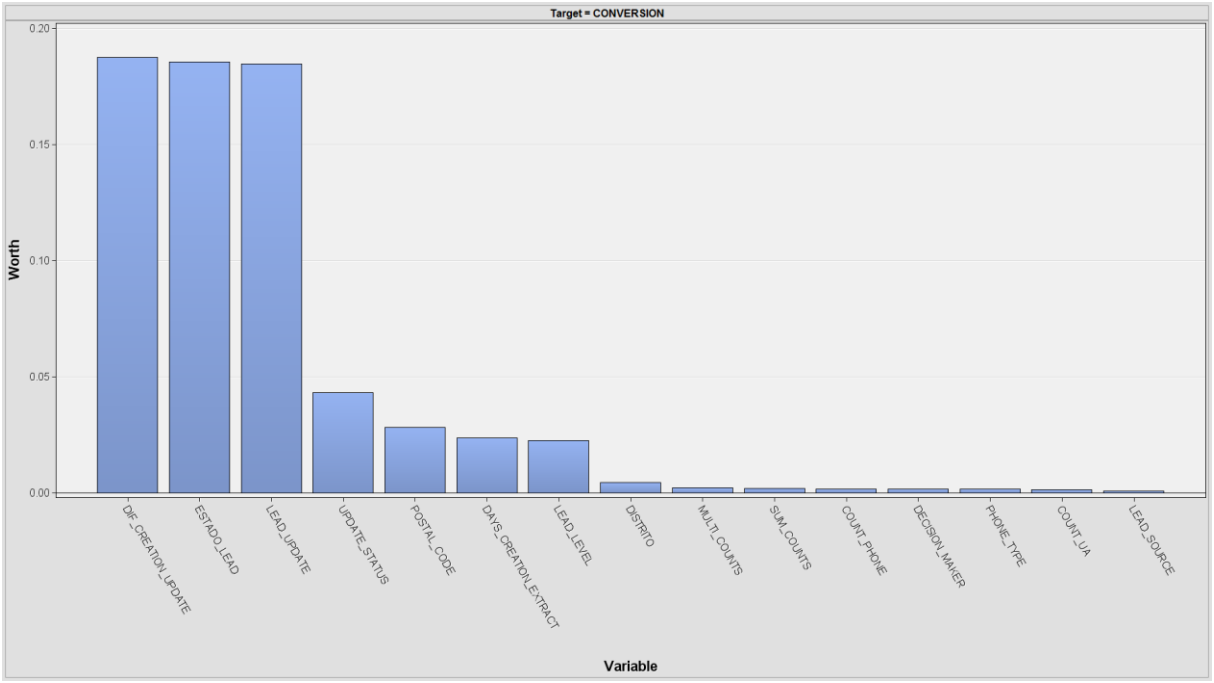

Figura 6.17 - *SAS Enterprise Miner* - Representação do valor de cada variável face à variável dependente.

Finalizado o tratamento e limpeza dos dados, prosseguiu-se para a componente de partição dos dados com um conjunto de 15 variáveis e 103 422 registos.

6.4.2.3. Partição dos Dados

Na secção 4.3.3 foram abordados os três tipos de conjuntos de dados necessários para a construção de um modelo de preditivo. Esta partição dos dados garante que o modelo é validado e testado com dados complementares aos que foram utilizados durante a fase de aprendizagem. Na construção do modelo de propensão de *leads* foi aplicado um componente disponibilizado pelo *software*, no qual foi necessário parametrizar a proporção de cada conjunto e selecionar o método de partição.

De acordo com as recomendações de Hand, Mannila e Smyth (2001), foram definidas as proporções 60% - 20% - 20% para os dados de treino, validação e teste, respetivamente. Foi selecionado o método de partição estratificado, o qual assegura a proporção da variável dependente nos vários conjuntos, o que tem um impacto positivo na precisão da classificação final. Na figura 6.18 é possível observar como foi efetuada a divisão. Repare-se que no conjunto original a proporção de classificações de sucesso e não sucesso é 56,7% e 43,3%, respetivamente. Como se pode verificar na figura, esta proporção é preservada nos conjuntos de treino, validação e teste.

Summary Statistics for Class Targets

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
CONVERSION	0	0	44758	43.2771	
CONVERSION	1	1	58664	56.7229	

Data=TEST

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
CONVERSION	0	0	8952	43.2756	
CONVERSION	1	1	11734	56.7244	

Data=TRAIN

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
CONVERSION	0	0	26854	43.2766	
CONVERSION	1	1	35198	56.7234	

Data=VALIDATE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
CONVERSION	0	0	8952	43.2798	
CONVERSION	1	1	11732	56.7202	

Figura 6.18 - *SAS Enterprise Miner* - Partição dos dados nos conjuntos de treino, validação e teste.

6.4.2.4. Seleção das Variáveis

Avaliou-se anteriormente que, apesar dos esforços de limpeza e tratamento dos dados, o risco de *overfitting* continua presente em muitas técnicas de *data mining*. Como foi referido no capítulo 4.4.3, este risco é potenciado com o aumento do número de variáveis de entrada do modelo. Um número elevado de variáveis não só aumenta o custo e tempo de computação, assim como prejudica a capacidade preditiva do modelo. Deste modo, uma das tarefas mais importantes passa por selecionar criteriosamente as melhores variáveis, definindo um conjunto de dados reduzido, com a máxima relevância e menor redundância possível.

Na componente prática do estudo foram utilizados três métodos de seleção de variáveis, com o objetivo de identificar aquele que produz melhores resultados. Os métodos aplicados, e descritos no capítulo 4.4.3 referente à redução de dimensionalidade, foram os seguintes:

- Regressão Linear Múltipla (Seleção *Stepwise*)
- Componentes Principais
- Módulo "*Variable Selection"*

O método de Regressão *Stepwise*, enquanto método de seleção iterativo, inicia a análise sem potenciais variáveis candidatas para o modelo e de seguida, avalia iterativamente se o efeito de cada variável está significativamente associado com a variável dependente. O processo de adição ou remoção continua até uma das condições de convergência se verificarem: mais nenhum efeito admite o nível de significância mínimo ou é atingido número máximo de iterações definido.

Na aplicação prática deste método, definiu-se o nível mínimo de significância igual a 0.0001 e obteve-se um total de 13 iterações, descritas na tabela 6.5.

Tabela 6.5 - Processo de seleção de variáveis com o método de Regressão (Stepwise).

Iteração	Descrição da ação executada
1	Adição da variável Estado_Lead
2	Adição da variável Update_Status
3	Adição da variável Dif_Creation_Update
4	Adição da variável Postal_Code
5	Adição da variável Lead_Level
6	Adição da variável Decision_Maker
7	Adição da variável Count_Phone
8	Adição da variável Count_UA
9	Adição da variável Phone_Type
10	Adição da variável Lead_Update
11	Remoção da variável Phone_Type
12	Adição da variável Phone_Type
13	Remoção da variável Phone_Type

Deste modo, como resultado da seleção efetuada pelo método de Regressão *Stepwise*, segue para a fase de modelação um subconjunto de dados de treino, validação e teste com 9 atributos.

A segunda abordagem adotada como método de seleção de variáveis foi a análise de componentes principais. Perante a transformação das variáveis de entrada em componentes principais, assume-se que, apesar de ser um método bastante eficaz na redução da dimensionalidade, dificulta a interpretação final

dos resultados. No desenvolvimento deste método foi utilizado componente PCA (*Principal Components Analysis*) com a parametrização apresentada na figura 6.19.

Property	Value
General	
Node ID	PRINCOMP
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Eigenvalue Source	Correlation
Interactive Selection	
Print Eigenvalue Source	No
Score	
Princomp Prefix	PC
Eigenvalue Cutoff	
Cumulative	0.99
Increment	0.001
Max Number Cutoff	
Apply Maximum Number	Yes
Maximum Number	10
Reject Original Input Variables	Yes
Hide Rejected Variables	Yes

Figura 6.19 - *SAS Enterprise Miner* - Configuração do módulo de Componentes Principais.

Conforme apresentado, os valores transformados foram calculados com base na matriz de correlação das variáveis, e foram definidos os critérios de paragem: variância total acumulada máxima igual a 0,99 ou um incremento de valor mínimo de 0,001. Inicialmente não foi definido um número máximo de componentes principais, porém, recorreu-se à função de seleção interativa para avaliar o decréscimo de valor de componente para componente. Esta funcionalidade permite realizar uma seleção visual do número ideal de componente principais que deverão resultar deste processo. Na figura 6.20 é ilustrado o valor acrescentado por componente quando adicionado ao conjunto.

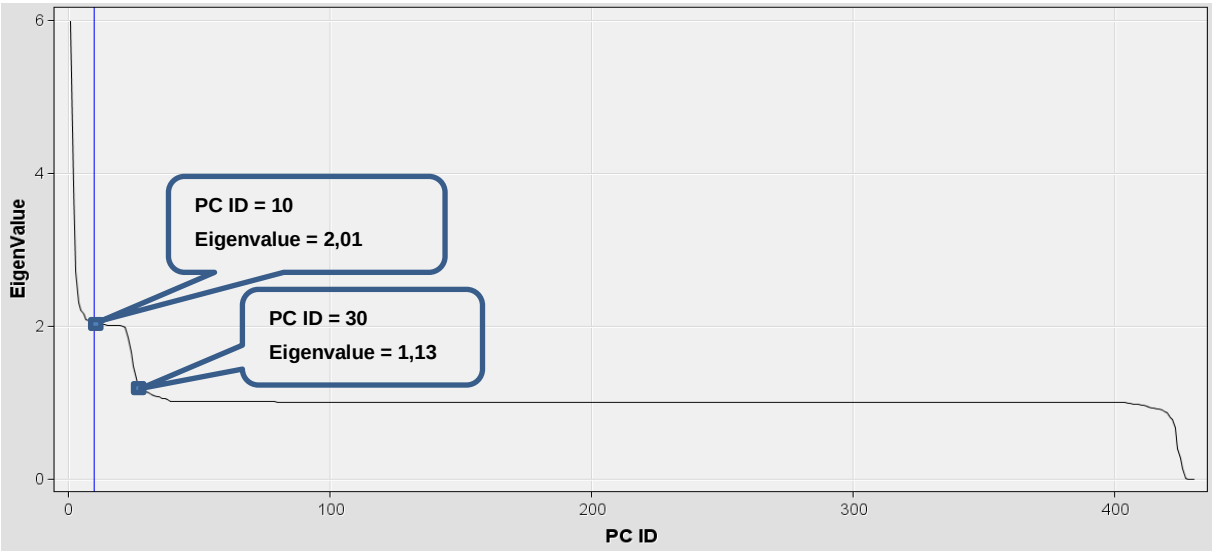

Figura 6.20 - *SAS Enterprise Miner* - Seleção interativa do número de componentes principais.

Assim, o objetivo passa por identificar o ponto em que o decaimento de valor passa a ser mais "suave", podendo descartar aqueles que apresentam menor contribuição. Observando a figura 6.20, identificam-se duas zonas onde a taxa de decrescimento é menor, que correspondem a 10 ou 30 componentes principais. Como o objetivo deste processo é reduzir efetivamente a dimensionalidade do modelo, optou-se por escolher os 10 primeiros componentes. Na figura 6.21 verifica-se que com os primeiros 10 componentes principais é possível explicar 6,03% da variância total das variáveis. Este valor é considerado baixo, mas é potencialmente provocado pela variável código postal que introduz muita dispersão dos dados.

Eigenvalues of Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	5.98665053	2.12643538	0.0131	0.0131
2	3.86021515	1.14347988	0.0085	0.0216
3	2.71673527	0.41053913	0.0060	0.0276
4	2.30619614	0.09713993	0.0051	0.0326
5	2.20905621	0.05953907	0.0048	0.0375
6	2.14951714	0.07073683	0.0047	0.0422
7	2.07878031	0.01154591	0.0046	0.0467
8	2.06723440	0.01108192	0.0045	0.0513
9	2.05615247	0.00744809	0.0045	0.0558
10	2.04870438	0.00885257	0.0045	0.0603

Figura 6.21 - *SAS Enterprise Miner* - Estatísticas associadas ao 10 primeiros Componentes Pricipais.

Por fim, o último método consistiu na aplicação do módulo de *Variable Selection*. Este módulo identifica as variáveis com um "bom poder preditivo" face à variável dependente e seleciona-as como entrada para os módulos procedentes. O critério de avaliação definido baseia-se na métrica R-Quadrado, tendo sido configurados os seguintes valores: valor mínimo de R-Quadrado igual a 0,005 e limite mínimo de contribuição de R-Quadrado para o modelo igual a 0,0005. Analisando os resultados obtidos, apresentados na figura 6.22, verifica-se que foram selecionadas 5 variáveis, sendo as restantes rejeitadas por apresentarem um R-Quadrado inferior ao estabelecido.

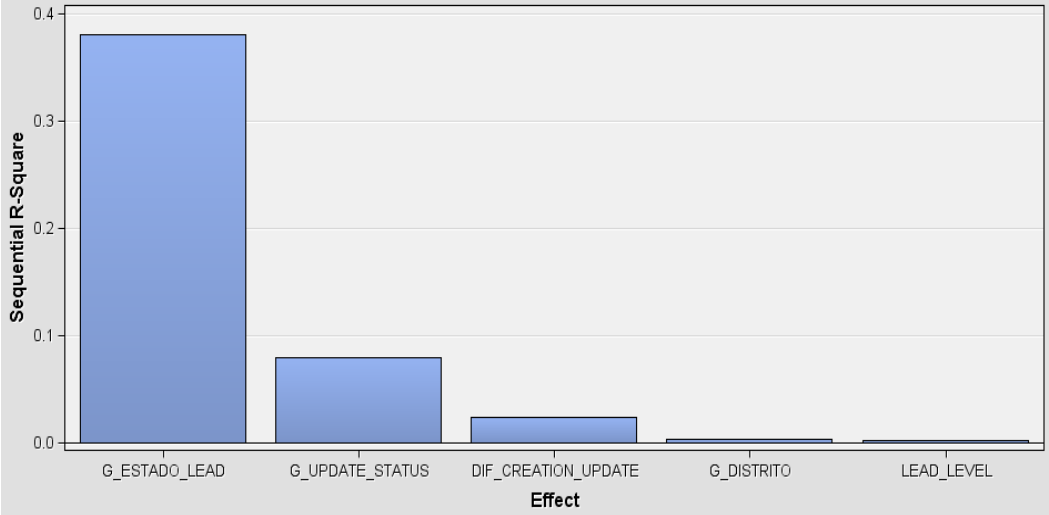

Figura 6.22 - *SAS Enterprise Miner* - Valor de R-Quadrado introduzido por cada variável selecionada.

Esta seleção garante uma precisão de 83%, aproximadamente a mesma obtida pelo método de Regressão. Deste modo, como resultado da seleção efetuada pelo método de *Variable Selection*, segue para a fase de modelação um subconjunto de dados de treino, validação e teste com 5 variáveis.

6.4.3. Modelação dos Dados

Após a execução da redução de dimensionalidade procedeu-se a fase de modelação, onde as variáveis selecionadas serviram de entrada para os modelos escolhidos. Para cada método de seleção de variáveis foram aplicados 11 modelos distintos, conforme apresentado na figura 6.23.

Nesta fase cada modelo é treinado com o conjunto de dados de treino, a aprendizagem é controlada e afinada através do conjunto de validação e, por fim, o desempenho de cada modelo é avaliado recorrendo aos dados de teste. O módulo de Comparação de Modelos permite comprar as estatísticas para os vários modelos. A figura 6.23 é ilustrativa dos modelos aplicados, sendo eles redes neuronais, árvores de decisão, *gradient boosting*, regressão e *ensemble.*

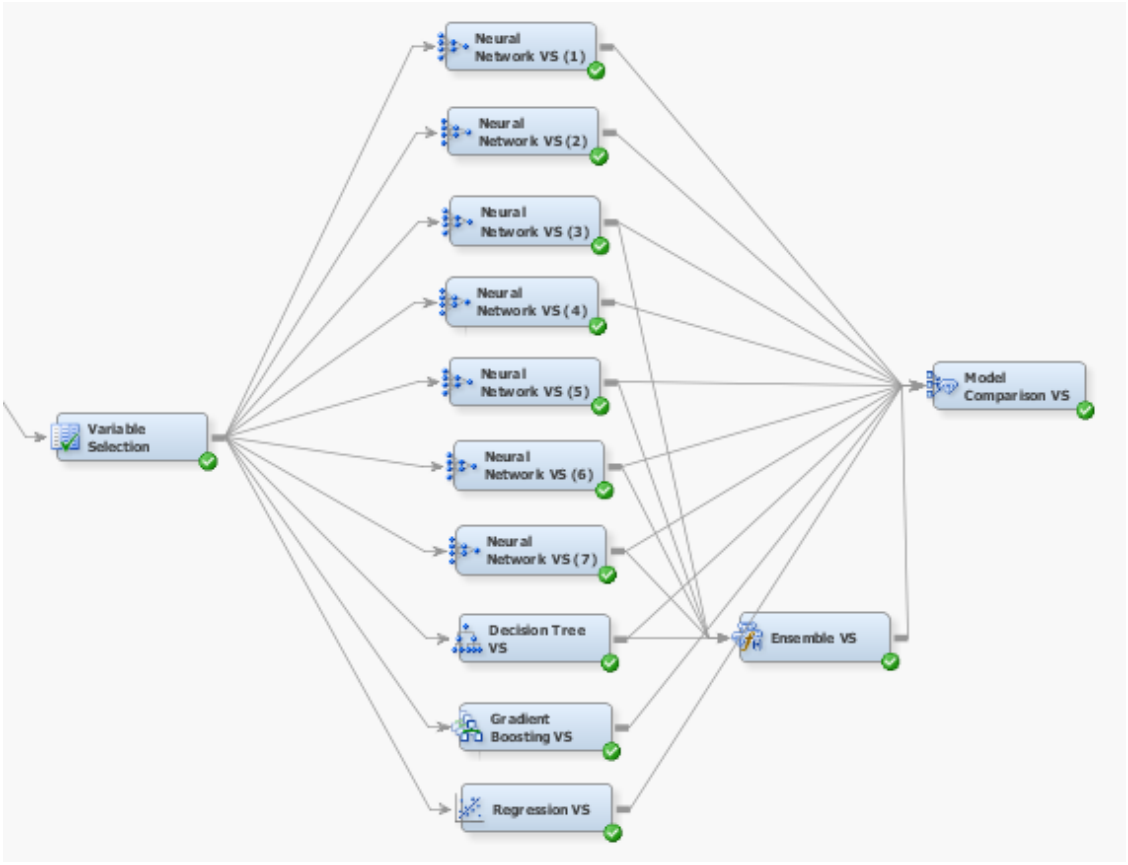

Figura 6.23 - *SAS Enterprise Miner* - Representação da fase de comparação de modelos.

No total foram comparadas 7 redes neuronais, diferindo entre elas no número de neurónios presentes na camada escondida, variando entre 1 e 7. Relativamente às árvores de decisão foi definido que, de modo a evitar a sobreaprendizagem da árvore, o número máximo de divisões por nó e a profundidade máxima da árvore seriam igual a 3, conforme a parametrização refletida na figura 6.24. Na definição da medida de avaliação da qualidade de partição de cada nó optou-se pela entropia, tal como recomendado no capítulo 4.5.2. Os restantes parâmetros apresentados na figura 6.24 mantiveram os valores por defeito aplicado pelo módulo.

Train	
Variables	
Interactive	
Import Tree Model	No
Tree Model Data Set	
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	Yes
Maximum Branch	3
Maximum Depth	3
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25

Figura 6.24 - *SAS Enterprise Miner - Parametrização do módulo Árvore de Decisão.*

Na configuração do modelo *Gradient Boostig* manteve-se os mesmos parâmetros para a sequência de árvores de decisão, especificando um número máximo de modelos produzidos igual a 50 como critério de paragem. Por sua vez, para o modelo de regressão foi selecionada a opção de regressão logarítmica por ser a mais adequada para a classificação de variáveis binárias, para um conjunto de dados não linear.

Por fim, foi aplicado o método de *ensemble* com o objetivo de combinar o poder preditivo dos vários modelos e assim, obter um modelo melhorado e com maior precisão na classificação de novos objetos. A estratégia de combinação de classificações adotada foi o valor médio das probabilidades posteriormente calculadas, considerando os 4 modelos com melhor desempenho nos dados de teste.

Na seção seguinte apresentam-se os resultados de cada modelo, assim como a comparação das métricas que permitem avaliar a qualidade e precisão da classificação.

6.4.4. Comparação dos Resultados e Classificação

Na comparação de modelos podem ser adotadas várias abordagens de modo a avaliar a qualidade dos modelos e selecionar aquele com melhor desempenho. Se por um lado é relevante analisar as métricas

que caracterizam a capacidade preditiva do modelo, por outro é também importante considerar os objetivos de negócio e a rentabilidade de campanhas futuras. Na comparação estatística dos modelos, apesar da ferramenta calcular inúmeras estatísticas, optou-se por analisar as seguintes métricas de desempenho: taxa de erro (*missclassification rate)*, coeficiente de Gini, índice ROC e a estatística de Kolmogorov-Smirnov. Note-se que dificilmente existe um modelo vencedor para todas as métricas, sendo por isso utilizados critérios de negócio como desempate.

A taxa de erro representa a percentagem de registos classificados incorretamente, ou seja, a percentagem de falsos positivos e falsos negativos no total de eventos. O coeficiente de Gini mede a uniformidade da distribuição, variando entre 0 e 1. O índice ROC, como referido anteriormente, representa a área debaixo da curva ROC, que varia numa escala de 50% (para modelos aleatórios) a 100%. A estatística de Kolmogorov-Smirnov representa a distância máxima entre a distribuição dos eventos e dos não eventos. Graficamente, trata-se da distância vertical máxima entre a linha de Sensibilidade e a base da curva ROC, ou seja 1- Especificidade, o que justifica a alta correlação com os valores da estatística da curva ROC. Resumidamente, o objetivo é selecionar o modelo com a menor taxa de erro, e maior valor para o coeficiente de Gini, índice ROC e estatística de Kolmogorov-Smirnov para o conjunto de dados de teste. Outro indicador bastante intuitivo é a representação gráfica das curvas ROC. A curva de cada modelo exibe um grau de concavidade diferente, sendo que quanto maior a concavidade e quanto mais próximo estiver do canto superior esquerdo, melhores são os resultados esperados. As figuras A.1, A.2 e A.3, presentes no anexo A, mostram as curvas ROC dos diferentes modelos, para os vários métodos de seleção de variáveis, Regressão, PCA e *Variable Selection*, respetivamente.

Deste modo, prosseguiu-se à comparação das estatísticas referidas para cada um dos métodos de seleção de variáveis, considerando os resultados para o conjunto de dados de teste. Nas tabelas em anexo, presentes no anexo B, apresentam-se os valores ordenados do melhor para o pior, consoante cada estatística. Apesar de existirem valores muito próximos, pôde concluir-se que, na generalidade, os modelos de *Ensemble* são aqueles com os melhores resultados, pois combinam as capacidades preditivas dos melhores modelos. Como se pode verificar nas tabelas B.1, B.2 e B.3 os modelos de *Ensemble* ocupam quase sempre os primeiros lugares na tabela.

Por fim, interessa avaliar dos 3 modelos produzidos, aquele que é o vencedor absoluto. Na tabela 6.6 apresenta-se um resumo dos valores das estatísticas dos modelos de *Ensemble* obtidos por cada um dos métodos de seleção.

Tabela 6.6 - Comparação dos modelos de Ensemble produzido pelos diferentes métodos de seleção de variáveis.

Modelo	Taxa de Erro	Índice ROC	Coeficiente de Gini	Estatística de Kolmogorov-Smirnov
Ensemble Regressão	0,1618	0,894	0,788	0,710
Ensemble PCA	0,1644	0,890	0,780	0,705
Ensemble VS	0,1665	0,878	0,755	0,702

Comparando os valores para as 4 métricas, pôde concluir-se que o método de regressão foi o que produziu um modelo Ensemble mais forte, seguido do método de análise de componentes principais (PCA), e por fim, o método automático de *Variable Selection.*

Assim, identificou-se o modelo *Ensemble*, obtido pela junção das redes neuronais com 2, 3, 5 e 7 neurónios na camada escondida e resultante da seleção de 9 variáveis pelo método de Regressão (*Stepwise)*, como o modelo com melhor desempenho e maior potencial para a classificação de novas *leads.* Posto isto, avançou-se para a fase de classificação considerando exclusivamente este modelo.

Na fase de classificação foi calculada a probabilidade de conversão de cada novo registo introduzido no modelo e estimada a respetiva previsão de classificação. A figura 6.25 demonstra a fase de importação das novas *leads* e consequente classificação, segundo o modelo selecionado, neste caso o modelo *Ensemble.*

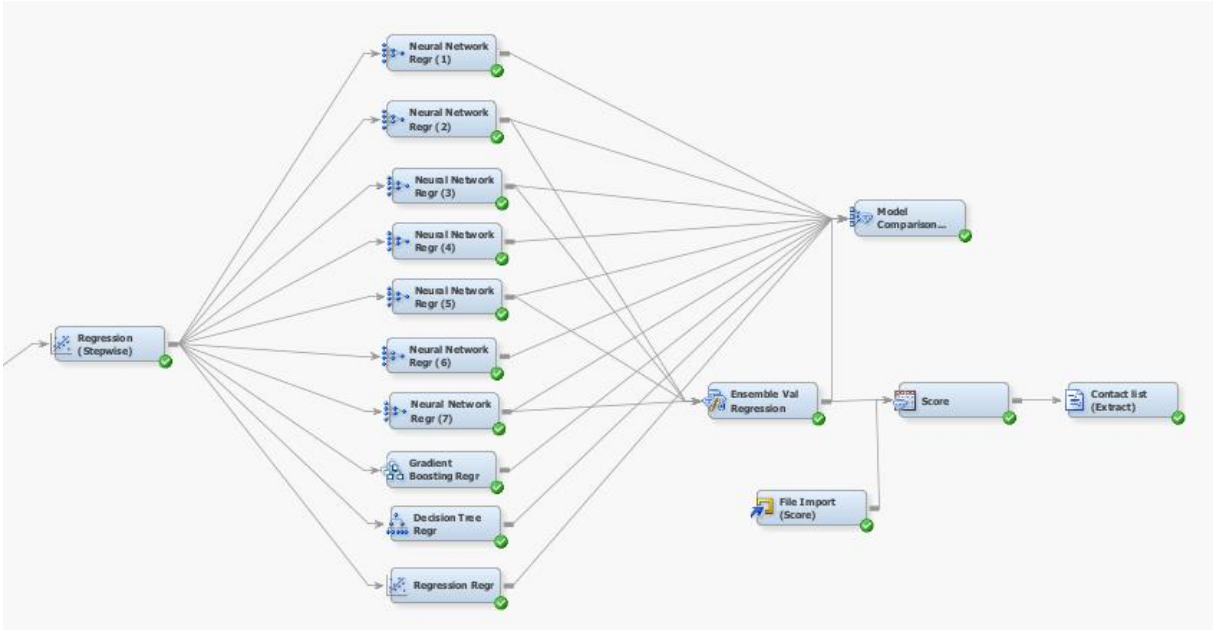

Figura 6.25 - *SAS Enterprise Miner* - Classificação de novos registos com o modelo *Ensemble*, obtido pelo método de seleção Regressão (*stepwise*).

Por fim, criou-se um código que permitiu extrair um ficheiro Excel com o ID das *leads* e respetiva previsão probabilidade de conversão, ordenado por ordem decrescente de probabilidade. Atendendo às restrições de negócio mencionadas anteriormente, foi pré-definido um total de 85 000 *leads*, que corresponde ao número máximo de contactos possíveis de realizar num mês. A figura 6.26 exibe uma amostra das 15 primeiras linhas do ficheiro extraído.

	A	B
1	E_LEAD_ID	P_CONVERSION1
2	800038512	0,988671228
3	800062925	0,988671228
4	600144208	0,987885424
5	700108599	0,981416121
6	800105240	0,981416121
7	300031871	0,975290993
8	400031189	0,975290993
9	700051870	0,975290993
10	200006936	0,960426487
11	200115646	0,960426487
12	300008723	0,960426487
13	300009629	0,960426487
14	500026345	0,960426487
15	600109412	0,960426487

Figura 6.26 - Lista de *leads* e respetiva probabilidade de conversão, ordenada por ordem decrescente.

De modo a caracterizar melhor o conjunto de dados obtido, agrupou-se as *leads* em classes e produziu-se o seguinte histograma, apresentado na figura 6.27. Na tabela 6.7 apresenta-se a frequência absoluta de *leads* em cada classe.

Tabela 6.7 - Distribuição de *leads* por probabilidade de conversão, em classes.

Classes	Frequência Absoluta	Frequência Acumulada	Percentagem Acumulada
[90% - 100%]	260	260	0,31%
[80% - 90%[	31	291	0,34%
[70% - 80%[	1	292	0,34%
[60% - 70%[	189	481	0,57%
[50% - 60%[	27 322	27 803	32,71%
[40% - 50%[	19 687	47 490	55,87%
[30% - 40%[	37 510	85 000	100%
[0% - 30%[	0	85 000	100%

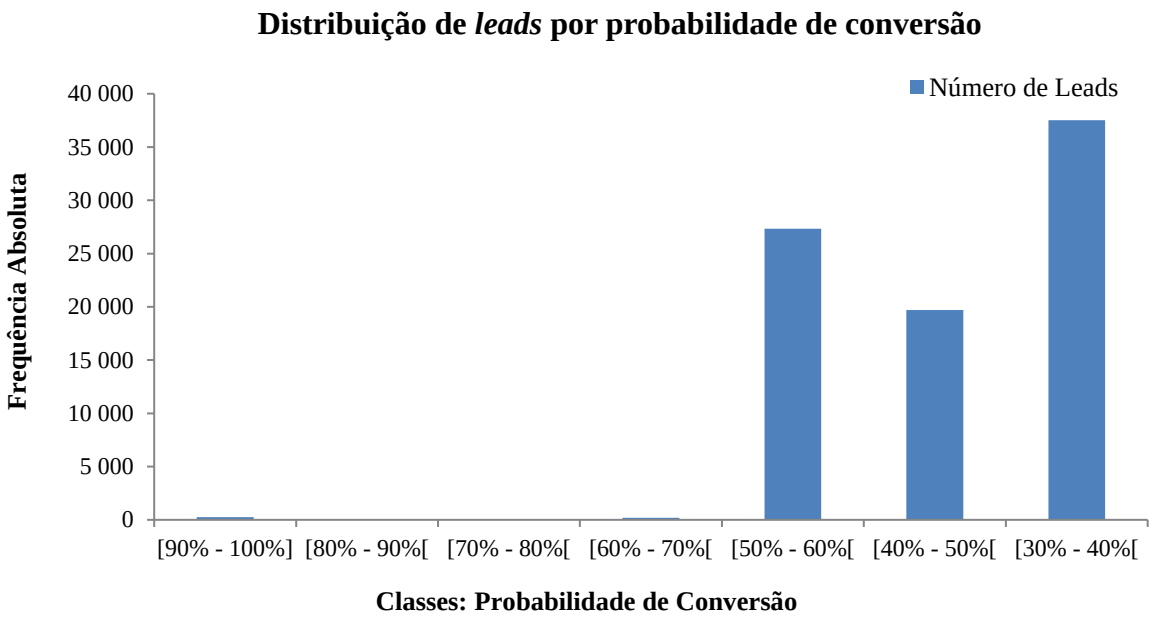

Figura 6.27 – Distribuição das *leads* extraídas, agrupadas em classes de probabilidade de conversão.

Como se pôde verificar, através das figuras acima representadas, a 67.3% das *leads* encontram-se entre os 30% e os 50% de probabilidade de conversão, sendo que apenas 0,57% tem uma probabilidade superior a 70%.

Assim, foi atingido o objetivo de identificar as *leads* mais propensas para a conversão de um serviço fixo e criado o mecanismo de extração da lista a ser entregue aos parceiros de *telemarketing* com os contactos a efetuar. A estratégia de contactos fica a cabo da decisão da empresa, pois deve ser gerida consoante a disponibilidade e capacidade subcontratada dos parceiros em cada mês. Extraindo o máximo de contactos exequíveis, cumpre-se o propósito de apresentar os vários cenários possíveis com as respetivas vantagens e desvantagens inerentes. No capítulo seguinte apresentam-se as várias opções, decorrentes dos resultados obtidos, assim como as recomendações, dadas a taxas de conversão prevista e custos associados.

6.5. Recomendações de melhoria

Ao longo do capítulo 6 foi descrito o caso de estudo que suportou a investigação realizada. Numa primeira instância foram analisados os processos de captura, tratamento e conversão de *leads* implementados por uma empresa de telecomunicações, e com base nas dificuldades reportadas e problemas identificados na análise de dados capturado, foram enumeradas 11 oportunidades de melhoria. Os pontos referidos abrangem as diferentes componentes da gestão de *leads*, sendo que se procurou encontrar possíveis soluções, de modo a responder às necessidades da empresa.

De acordo com as prioridades descritas, a empresa alocou maior criticidade à melhoria da qualidade dos dados e à necessidade de uma ferramenta que facilitasse e tornasse mais eficaz o processo de segmentação. Nesse sentido, foi apresentado, nas secções 6.3 e 6.4, o desenvolvimento de um modelo de limpeza dos dados e um modelo preditivo da propensão das *leads*.

Neste subcapítulo propõe-se a apresentação e discussão das propostas de solução, quer para cada um dos pontos de melhoria identificados, quer para a segmentação final tendo por base os resultados do modelo preditivo.

6.5.1. Proposta de soluções para os problemas identificados

No seguimento dos problemas identificados no capítulo 6.2.1, apresentam-se de seguida um conjunto de recomendações que visam a resolução numa perspetiva a curto e médio-longo prazo. Idealmente, as medidas a curto prazo referem-se a alterações processuais ou de funcionamento entre canais, assistentes ou outros intervenientes, enquanto que a medidas a médio prazo implicam alterações de implementação, novos desenvolvimentos ou dependências com outras áreas.

Assim, recapitulando também os 11 aspetos apresentados anteriormente, seguem-se as sugestões de melhoria organizadas pelas respetivas etapas do processo de gestão de *leads*.

Captura

1- Reduzido número de *leads* produzidas pelos canais de vendas

Curto Prazo: Reforçar o procedimento de captura e atualização de *leads* qualificadas junto dos assistentes responsáveis pelos canais de vendas; Criar envolvimento e motivação através da definição de objetivos pessoais concretos.

2- Elevado número de *leads* com dados inválidos, incorretos ou inconsistentes

Curto Prazo: Estabelecer um procedimento que identifique todas as *leads* inválidas, com dados incorretos ou inconsistentes e forçar a alteração (em lote) do estado da *lead* para "Fechado". Deste modo, a *lead* deixa de ser considerada em análises e campanhas posteriores, evitando o desperdício de recursos; Aplicação recorrente do modelo de limpeza sugerido.

Longo Prazo: Criar mecanismos de validação de informação no website, mais precisos e rigorosos. Por exemplo, bloquear a introdução de números de telefone falsos e frequentemente utilizados, normalizar os campos de morada e apresentar resultados pré-definidos para a morada, mediante o código postal indicado.

3- Duplicação de *leads* no mesmo canal

Curto Prazo: Aplicação recorrente do modelo de limpeza sugerido.

Longo Prazo: Para as *leads* capturadas através do *website,* criar um mecanismo que evite o envio de informação de registos consecutivos com a mesma informação, por exemplo, aplicada aos campos do número de telefone e morada. Adicionalmente, criar um indicador na base de dados que aumentasse incrementalmente sempre que aquela combinação de valores fosse solicitada. Deste modo, mantinha-se o registo do número de vezes que a *lead* era pesquisada, e era possível atualizar continuamente a data de última atualização. Relativamente à captura em canais assistidos, sugere-se a implementação de uma lógica na interface de utilizador que, no momento de guardar uma nova *lead,* identifique se existe na base de dados alguma *lead* com a mesma informação, e dê ao assistente a possibilidade de atualizar a *lead* em vez de criar uma nova.

Enriquecer

4- Duplicação de *leads* para o mesmo utilizador

Curto Prazo: Aplicação recorrente do modelo de limpeza sugerido. Aplicação do modelo de propensão, garantindo a seleção da *lead* e do canal com maior valor.

Longo Prazo: Os mecanismos referidos no ponto 3 aplicam-se também como recomendações e possíveis soluções para resolver o ponto 4.

5- 90% das *leads* são incompletas

Curto Prazo: Reforçar o procedimento de enriquecimento e introdução dos dados no *software* junto dos assistentes, para que seja coletada e registada mais informação em todos os pontos de contacto com potenciais cliente; criar iniciativas de nutrição e enriquecimento, incentivando as *leads* a fornecerem informação através do envio de email, SMS ou chamada.

Longo Prazo: Todos os mecanismos já referidos que reduzem a duplicação de *leads*, e consequentemente, condensam toda a informação numa única *lead* mais válida e qualificada; tornar o preenchimento da data de final de contrato (CED) obrigatório em canais assistidos;

6- Desatualização do campo "Elegibilidade"

Curto Prazo: Adotar um procedimento com a validação da elegibilidade no final do modelo de propensão, para que não sejam excluídas *leads* cujo campo está desatualizado. Uma forma de simplificar o processo para o imediato é cruzando as duas informações manualmente, através de simples *queries* de SQL.

Longo Prazo: Adotar um procedimento de validação de elegibilidade automático, capaz de atualizar continuamente os valores inseridos na base de dados, caso existam alterações nas zonas de cobertura.

Nutrição

7- Falta de processo de acompanhamento para *leads* qualificadas

Curto Prazo: Criar iniciativas de nutrição das *leads*, promovendo o compromisso e a relação com a marca. Quer via SMS, email ou por chamada, a oferta e abordagem pode ser diferenciada com base na informação coletada. Por exemplo, uma *lead* entre os 6 e os 3 meses do final do contrato requer uma mensagem mais direta e comercial do que outra a mais de 6 meses.

Rastrear

8- Atualização manual do estado

Curto Prazo: Reforçar o procedimento atualização de *leads* e alteração do estado junto dos assistentes responsáveis pelos canais de vendas; **e**stabelecer um procedimento que identifique todas as *leads* inválidas, com dados incorretos, duplicadas, ou para as quais já existiu conversão, embora não tenha ficado registado, e forçar a alteração (em lote) do estado da *lead* para "Fechado". Deste modo, a *lead* deixa de ser considerada em análises e campanhas posteriores, evitando o desperdício de recursos;

Longo Prazo: Implementar um novo fluxo de estados das *leads*, mais intuitivo e completo, que atualize automaticamente sem necessidade da intervenção manual dos assistentes, com base em regras de negócio.

Avaliar

9- Falta de métricas de desempenho

Curto Prazo: Instituir procedimentos de medição e avaliação das campanhas, através do registo rigoroso do desfecho de cada chamada efetuada no âmbito de uma campanha. Calcular a taxa de conversão com base nas adesões efetivas.

10- Sem procedimentos de auditoria a avaliação da qualidade dos dados.

Curto Prazo: Instituir um indicador para medir a qualidade e confiança nos dados produzidos, por canal. Comparar resultados após aplicações consecutivas do modelo de limpeza, de modo a avaliar a melhoria conquistada e a identificar situações em será necessário afinar os critérios do modelo.

11- Falta de prioritização das *leads* com base no seu valor

Curto Prazo: Desenvolvimento de um modelo de propensão que, com base em comportamentos anteriores, é capaz de aprender padrões e regras escondidas nos dados capazes de prever o sucesso ou insucesso de uma conversão. Deste modo, é calculada a probabilidade de conversão de cada *lead*, e são selecionadas aquelas com maior propensão para aderir a determinada oferta.

6.5.2. Proposta de soluções para a segmentação de *leads*

Concluído o desenvolvimento do modelo de propensão, que permitiu a atribuição de uma probabilidade de conversão a cada uma das *leads* do conjunto de dados, e avaliados os resultados obtidos, tornou-se viável a formulação de possíveis soluções para a segmentação da próxima campanha. Na formulação de hipóteses, é importante considerar as restrições e regras impostas pelo negócio. Nesse sentido, e atendendo às condicionantes relativas à alocação de parceiros e respetivas características, foram identificadas um total de 7 possíveis ações, apresentadas na tabela 6.8.

Tabela 6.8 - Conjunto de soluções finais para a segmentação do conjunto de *leads.*

Soluções	Parceiros de Telemarketing	Número Máximo de Contactos (contactos/mês)	Custo Unitário (€/contacto)	Custo Total (€)	Taxa de Sucesso Esperada (%)	Prioridade
1	Parceiro A	15 000	3	45 000	57,5%	1º
2	Parceiro B	25 000	4	100 000	57,2%	2º
3	Parceiro C	45 000	6	270 000	52,0%	4º
4	Parceiro A e B	40 000	3,6	145 000	53,3%	3º
5	Parceiro A e C	60 000	5,3	315 000	48,7%	5º
6	Parceiro B e C	70 000	5,3	370 000	46,9%	6º
7	Parceiro A e B e C	85 000	4,9	415 000	44,6%	7º

As soluções apresentadas na tabela 6.8 abrangem a utilização de 1, 2 ou os 3 parceiros de telemarketing. O Parceiro A, sendo o que pratica o menor custo unitário e o menor número de contactos, ostenta também ser o mais o vantajoso caso se pretenda fazer um número reduzido de contactos. O Parceiro B, por sua vez, permite efetuar mais 10 000 contactos face ao A, com uma taxa de conversão semelhante e com um custo ligeiramente maior. Comparando as soluções 3 e 4, torna-se evidente que seria mais vantajoso subcontratar os dois parceiros em vez do parceiro C, pois este praticar um custo unitário bastante mais elevado. A solução 7 é aquela que procura utilizar a totalidade dos recursos, assumindo uma taxa de sucesso esperada mais baixa e implicando o maior custo por mês.

Com base nesta tabela, a empresa tem na sua alçada o poder e a fundamentação para a tomada de decisão. Quer o critério mais relevante seja o custo total, quer seja a taxa de sucesso exigida, esta tabela permite considerar todos os cenários possíveis, e adotar aquele que melhor se adequa às necessidades momentâneas da empresa, funcionando como uma ferramenta de suporte.

Do ponto de vista analítico, foram ponderadas as recomendações a adotar, as quais foram expressas na coluna Prioridade da tabela 6.8. Note-se que a sugestão foi calculada face aos resultados obtidos pelo modelo de previsão, e que poderá sofrer alterações de mês para mês, consoante a propensão das *leads* extraídas e mediante alterações às condições impostas pelos parceiros. Para além disso, a recomendação deverá ser sempre avaliada à luz da estratégia da empresa no corrente mês, justificando um investimento maior ou menor no número de contactos a realizar. Assim, sugere-se como solução prioritária a subcontratação do Parceiro A, pois é o que garante uma taxa de sucesso maior, com um menor custo total.

7. Conclusões e Limitações

7.1. Conclusões

O presente trabalho de investigação surge da necessidade real, identificada por empresas de diversos setores, de um modelo preditivo que suporte a tomada de decisão no processo de segmentação de *leads*, garantindo que são selecionadas as *leads* corretas, na quantidade correta e aumentando significativamente a taxa de conversão esperada. Assim, foram formuladas as seguintes questões para abordar o problema:

1. Como é que as técnicas de *data mining* podem auxiliar a tomada de decisão no processo de segmentação de *leads*?
2. Como aumentar a taxa de conversão de *leads*?

Atendendo a ambas as questões, foram traçados cinco objetivos, os quais foram alcançados ao longo da dissertação. O primeiro objetivo estava relacionado com a compreensão do fluxo de processos na gestão de *leads*, desde a captura, ao tratamento e conversão. Neste âmbito foi conduzida uma revisão do estado da arte na literatura disponível, e avaliado o trabalho desenvolvido nesta área pelos diferentes autores. Embora exista um maior foco na eficácia dos processos de conversão, visto ser o derradeiro objetivo das empresas, há um consenso relativamente à integração de um modelo de procedimentos que visam suportar e melhorar os resultados na fase de conversão. O modelo sugerido engloba 6 etapas, as quais destacam a captura, enriquecimento, nutrição, rastreamento, avaliação e conversão das *leads*. Assim, através das ações de enriquecimento, nutrição e avaliação contínua dos desenvolvimentos de cada *lead*, é reforçada a qualidade e a maturidade da *lead*, potenciando que esta alcance a fase de conversão com maior probabilidade de sucesso.

O segundo objetivo definido incluía a identificação de melhorias nos processos, a curto e médio-longo prazo, com a finalidade de melhorar a qualidade dos dados introduzidos nas bases de dados. Este processo envolveu uma análise criteriosa dos procedimentos dos vários canais intervenientes, das dependências e relações entre canais, e das estratégias de marketing em vigor. Só assim foi possível identificar possíveis aspetos de melhoria e propor um plano de ação faseado. Os problemas identificados incidiam essencialmente na qualidade dos dados, quer pela introdução de dados incorretos e incoerentes, quer pela duplicação de *leads* com a mesma informação. A falta de investimento em ações de enriquecimento foi visível através da quantidade de *leads* pouco qualificadas e irrelevantes. Deste modo, foram sugeridos diversos pontos de ação, com o objetivo de melhorar os processos e os sistemas

tecnológicos envolvidos. Enumerando alguns dos mais relevantes, destacam-se o reforço de procedimentos junto dos assistentes dos canais de vendas, a introdução de iniciativas de enriquecimento e nutrição, e a automatização de processos de validação e atualização em todo o fluxo de dados. Deste modo, foram sugeridas no total 11 ações de melhoria, das quais 2 foram desenvolvidas no âmbito desta dissertação, correspondendo ao terceiro e quarto objetivo.

Assim, o terceiro objetivo compreendeu desenvolvimento de um modelo de limpeza, capaz de melhorar a qualidade dos dados armazenados. Numa perspetiva de redução dos dados redundantes e irrelevantes, procedeu-se à aplicação de técnicas de remoção de erros, de remoção de registos duplicados e de priorização, privilegiando os registos com maior valor para o negócio. O modelo foi construído na linguagem SQL, e resultou do encadeamento sequencial de 18 critérios, estabelecidos para remover instâncias repetidas e de fraca qualidade, potenciando as *leads* mais qualificadas e fiáveis. Como resultado, houve uma exclusão de 730 910 *leads*, o que representa uma redução de 57% do conjunto de dados original. A aplicação deste modelo, independentemente da existência de um modelo de propensão, garante que a segmentação incide sobre uma base de *leads* válida, com dados de contacto reais e com a informação mais atualizada.

De seguida, e satisfazendo o quarto objetivo proposto, foi desenvolvido um modelo de propensão através da aplicação de técnicas de *data mining* ao conjunto *leads* resultante do modelo de limpeza. De facto, também este tema foi alvo de uma revisão da literatura existente, como forma de fundamentação das técnicas e metodologias aplicada. No entanto, concluiu-se que, apesar de existirem extensas investigações referentes à aplicação de técnicas de *data mining* em casos reais de marketing direto para clientes, não foram identificadas obras referindo este mesmo estudo aplicado à gestão de *leads.* A identificação desta necessidade na literatura, motivou e conduziu o estudo efetuado à realidade da aquisição de clientes.

Com a finalidade de auxiliar a tomada de decisão no processo de segmentação de *leads,* foram utilizados dados reais, previamente classificados com eventos de sucesso ou insucesso, permitindo o treino de modelos preditivos. Através dos padrões e tendência inferidas dos dados, foi construído um modelo final, capaz de estimar a probabilidade de conversão de cada *leads.* A construção do modelo baseou-se nos procedimentos sugeridos pela metodologia SEMMA, começando pela exploração qualitativa e quantitativa dos dados. A fase de processamento foi a mais extensa e que exigiu maior esforço, incorporando as tarefas de limpeza, transformação das variáveis, partição dos dados e seleção das variáveis. Foram adotados e comparados os resultados de três métodos distintos de seleção de variáveis: Regressão (*stepwise*), Análise de Componentes Principais (PCA) e o módulo *Variable Selection.* De seguida, na fase de modelação foram comparados vários modelos como Redes Neuronais, Árvores de

Decisão, *Gradient Boosting*, Regressão e *Ensemble.* Por fim, a comparação e avaliação dos resultados ocorreu tanto ao nível dos modelos como dos métodos de seleção de variáveis, tendo sido escolhido aquele com melhores valores estatísticos nas diferentes métricas: taxa de erro, índice de ROC, coeficiente de Gini e estatística de Kolmogorov-Smirnov. Estas estatísticas expressam o poder preditivo de cada modelo. Assim, foi selecionado o modelo *Ensemble*, obtido pelo método de Regressão (*stepwise*). Este modelo resultou na combinação de vários modelos, beneficiando da capacidade preditiva e da variedade de classificações dos respetivos para a construção de um modelo mais completo e mais eficaz. De seguida, aplicou-se o modelo ao conjunto de dados produzido inicialmente, obtendo como resultado a classificação de cada *lead* e respetiva probabilidade. Ordenando por ordem decrescente, tornou-se fácil a seleção de *leads* com maior propensão para converter. Com estes desenvolvimentos, cumpre-se o objetivo de automatizar o processo de segmentação, suportando a seleção e atribuição eficiente de *leads.*

Por fim, o último objetivo expressava a capacidade de suportar a tomada de decisão na distribuição de contactos pelos diversos parceiros de telemarketing. Com este objetivo pretendia-se chegar a um conjunto de soluções possíveis de adotar pela empresa, com a possibilidade de serem avaliadas consoante a capacidade de recursos ou capacidade financeira disponível em cada mês. Analisando os dados produzidos pelo modelo de propensão, verificou-se que apenas 260 registos apresentavam uma probabilidade superior a 90%, e 292 com superior a 70%. Nesse sentido, formularam-se várias hipóteses com possíveis estratégias de ação. Essas hipóteses pretendem rentabilizar e tornar mais eficientes os contactos realizados em cada campanha, tendo em consideração as restrições e caraterísticas impostas pelo negócio. Conforme apresentado na tabela 6.8, a maior taxa de sucesso, cerca de 57,5%, é conseguida alocando 15 000 contactos ao Parceiro A. A utilização de todos os parceiros permite uma abrangência maior, 85 000 contactos por mês, no entanto, é esperada uma taxa de sucesso de 44,6%, representando o cenário de maior investimento para a empresa. Portanto, sugere-se uma segmentação de listas de contactos menores, uma vez que se consegue maior eficácia e qualidade, embora a decisão final deva ser ponderada pela empresa de acordo com a evolução das necessidades de negócio e respetivos requisitos.

O estudo das *leads* e do seu processo de criação, tratamento e conversão demonstrou-se fundamental para perceber a potencialidade dos dados recolhidos, mas também para identificar e justificar muitos dos erros e duplicações existentes nas bases de dados. De modo geral, pode afirmar-se que a investigação foi conduzida em duas vertentes: a primeira tinha como objetivo gerar dados com qualidade, isto é, atuar na fonte dos problemas e garantir que no futuro, com as sugestões de melhoria, esses problemas deixam de existir; a segunda tinha como objetivo potenciar a utilização dos dados já capturados e alocados na

base de dados, através da validação da qualidade e da segmentação mais eficaz e eficiente. A conceptualização das duas vertentes permite, assim, ter uma visão agregada do que deve ser um processo de gestão de *leads* para a empresa.

Mais especificamente no tratamento dos dados coletados, a combinação das técnicas de limpeza com estimativa de propensão para a conversão promete aumentar significativamente a eficiência e a rentabilidade das vendas por telemarketing. Juntos, estes dois mecanismos elevam a importância das *leads* nas empresas, contrariando a imagem de fracos resultados e fraca confiança nos dados, tornando as campanhas mais apelativas até para os assistentes de vendas.

No global, as questões de investigação foram respondidas com sucesso e foi demonstrado o grande potencial da informação escondida no enorme volume de dados capturados. Em particular, a criação de uma *lead*, para além de demonstrar um potencial interesse na aquisição de um serviço, revelou-se também uma valiosa fonte de informação para auxiliar as equipas de marketing na segmentação e tomada de decisão no lançamento de novas campanhas dirigidas a *leads*. Conclui-se que, quando corretamente aplicados nas estratégias de aquisição de novos clientes, os procedimentos descritos suportam uma alocação lógica e racional do orçamento e dos recursos disponíveis, operando como fator diferenciador na vantagem competitiva de qualquer empresa.

7.2. Limitações

O presente trabalho de investigação teve, no entanto, algumas limitações importantes de mencionar. Em primeiro lugar foi necessário garantir a proteção e confidencialidade dos dados, o que exigiu a anonimização e uma série de transformações das variáveis. Embora não se tenham perdidos dados, estas transformações podem ter impactado a deteção de padrões e a informação extraída. Também como forma de proteção, os dados foram extraídos numa data anterior à analise e à publicação do estudo, contudo, considera-se uma prova de conceito válida e apta para investigações futuras.

Relativamente ao modelo preditivo, houve claras limitações na qualidade dos dados disponibilizados. Como se observou no decorrer da dissertação, foram identificadas várias práticas que tinham um efeito prejudicial na qualidade dos dados, condicionando a qualidade do modelo e dos resultados finais. Para além das ações recomendadas neste âmbito, há ainda oportunidade de melhoria a nível dos sistemas tecnológicos e da arquitetura de informação que podem provocar um impacto positivo na qualidade do armazenamento, na rapidez da análise e da disponibilização dos resultados. No entanto, estas alterações dependem um investimento substancial em desenvolvimento de *software* e *hardware*.

Outro ponto de limitação no trabalho realizado centra-se na disponibilização da solução final. Enquanto que as propostas de melhoria e o modelo de limpeza, produzido em código SQL, puderam ser facilmente divulgados e partilhados, o modelo de propensão foi desenvolvido em ambiente local e alojado num único computador. As restrições de servidores e de ligações externas dificultou o processo de integração, impedindo a atualização dos dados em tempo real. Deste modo, o modelo carece sempre de uma extração dos dados, produzindo um atraso na utilização dos resultados. Por outro lado, estas restrições também não permitiram o desenvolvimento de um mecanismo de retrospeção, isto é, permitir que o modelo seja alimentado com os dados que produziu, para que aprenda com o sucesso ou com o erro de cada conversão estimada. Se houvesse uma integração em tempo real dos dados, os resultados de uma iteração poderiam ser incorporados no conjunto de treino da interação seguinte, enriquecendo e afinando o modelo a cada iteração.

A última limitação que carece de ser apresentada é a dificuldade em estimar a receita gerada por cada conversão. Idealmente, o plano de soluções proposto apresentaria uma ponderação entre os custos e a receita envolvida em cada ação. No entanto, a receita proveniente de uma aquisição de cliente é muito variável e diluída num período de tempo indeterminado. Pela especificidade e complexidade deste calculo optou-se por considerar os respetivos custos de subcontratação dos parceiros de telemarketing, por se tratarem de custos fixos.

Por fim, considera-se uma limitação ao trabalho desenvolvido o facto de não ter sido possível testar os resultados teóricos obtidos, através de uma campanha real. Tipicamente, há sempre um desvio associado entre os resultados teóricos e práticos, derivado de fatores de erro que não são considerados na teoria. Como validação dos modelos propostos, era importante realizar uma campanha piloto de modo a aferir o desempenho real e o desvio associado.

7.3. Recomendações de Trabalho Futuro

Neste último capítulo pretende-se apresentar recomendações para trabalhos futuros. Primeiramente, é importante mencionar que, apesar do caso de estudo ser centrado na gestão e conversão de *leads*, as técnicas e os procedimentos apresentados podem aplicados a qualquer conjunto de dados no âmbito de um problema de classificação.

Relativamente aos trabalhos de seguimento desta investigação, recomenda-se, em primeiro lugar, a alocação dos modelos desenvolvidos em servidores internos, integrados com as respetivas fontes de dados e respetivos sistemas de informação. Estes desenvolvimentos são imprescindíveis para a produção

de resultados atualizados e acionáveis em tempo real. Para além disso, a integração dos sistemas possibilitaria que o resultado de cada interação fosse automaticamente incorporado no conjunto de dados de treino, munindo o modelo de um mecanismo de avaliação e afinação constante. Como foi também mencionado na seção das limitações, futuramente, seria bastante interessante a criação de uma campanha de teste, de modo a testar a eficácia e o desempenho real dos modelos, e medir o impacto nas vendas relativamente a períodos anteriores.

Outro aspeto interessante seria complementar o estudo com informações e dados de clientes. Neste estudo de caso, a empresa tem um papel privilegiado de acesso a dados. Ora, pela exploração dos dados verificou-se que cerca de 44% das *leads* correspondem a atuais clientes móveis da operadora, existindo uma clara oportunidade de atuação. Para investigações futuras recomenda-se a análise isolada destas *leads,* complementando os dados com variáveis e características dos clientes, como idade, género, tarifário, consumo médio, antiguidade, entre outros. Estes são casos evidentes para os quais se pode adotar abordagens e ofertas personalizadas, existindo uma maior propensão para a conversão.

Logicamente, tanto o modelo de limpeza como o modelo de propensão foram desenhados e construídos à luz dos fatores e dos requisitos existentes no momento do estudo. No entanto, alterações a nível do negócio, da tecnologia utilizada, ou até atualizações aqui sugeridas podem tornar os modelos menos precisos ao longo do tempo. No modelo de limpeza, por exemplo, foram aplicados vários critérios baseados na experiência, no bom senso, e nas atividades de negócio. O modelo de previsão, por sua vez, baseou-se essencialmente em dados coletados *online*, os quais são bastante voláteis. São também esperadas mudanças resultantes das propostas de melhoria. Todos estes são fatores que causam impacto nos modelos e nos resultados produzidos, e, portanto, sugere-se uma revisão periódica com reavaliação dos critérios e afinação dos parâmetros ao longo do tempo.

Por fim, é apresentada uma recomendação baseada no estudo de Yang, Zhang e Zuo (2008). Como os autores indicam, o ser humano é um ser relacional, que age de acordo e é influenciado pela sua comunidade. Principalmente para o sector das telecomunicações, para o qual é claramente detetável a relação entre os vários clientes e possíveis clientes, faz sentido tirar partido dos atributos relacionais. No entanto, as empresas ainda não estão direcionadas para a recolha e tratamento de informação relacional, não existindo análises que contemplem o fluxo de comunicações, chamadas, mensagens, proximidade, entre outras características. Os autores propõem que estas características sejam utilizadas na construção de modelos de previsão, pois acreditam que estas ligações e relações de confiança têm um grande impacto nas decisões e comportamentos dos consumidores. Assim, seria interessante incluir todo este dinamismo em rede na análise de conversões de *leads*, pois são fatores que têm um claro impacto nos padrões de comportamento.

Referências Bibliográficas

Bairstow, T. (2016). Managing Online Lead Generation. *Pool & Spa Marketing, 40(3)*, 36-40.

Basheer, I. A., & Hajmeer, M. (2000). Artificial Neural Networks: Fundamentals, Computing, Design and Application. *Journal of Microbiological Methods, 43(1)*, 3-31.

Berry, L. (1995). Relationship Marketing of Services - Growing Interest, Emerging Perspectives. *Journal of the Academy of Marketing Science, 23 (4)*, 236 - 245.

Berry, L. (2002). Relationship Marketing of Services - Perspectives from 1983 and 2000. *Journal of Relationship Marketing, 1(1)*, 59 - 77.

Bhattacharyya, S. (1998). Direct Marketing Response Models Using Genetic Algorithms. In American Association for Artificial Intelligence (AAAI). Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (pp. 144 - 148). Chicago.

Blattberg, R., Kim, P., Kim, B., & Neslin, S. (2008). Acquiring Customers. In *Database Marketing: Analyzing and Managing Customers.* (pp. 495-514). London: Springer Verlag.

Buttle, F. (2009a). Introduction to customer relationship management. In *Customer relationship management: Concepts and technologies* (pp. 1-23). London: Taylor & Francis.

Buttle, F. (2009b). Managing the Customer Lifecycle: Customer Acquisition. In *Customer Relationship Management: Concepts and technologies* (pp. 225-254). London: Taylor & Francis.

Caller, R., Pallat, M., & Darlow, J. (2008). How do I ensure quality data from online lead generation? *Marketig Direct (1366896X)*, 42.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step data mining guide.* SPSS. CRISP-DM Consortium.

Chen, J.-S., & Ching, R. K. (2007). The Effects of Mobile Customer Relationship Management on Customer Loyalty: Brand Image Does Matter. Proceedings of the *40th Hawaii International Conference on System Sciences.* Hawaii: IEEE.

Chen, K., Kou, G., & Shang, J. (2014). An Analytic Decision Making Framework to Evaluate Multiple Marketing Channels. *Industrial Marketing Management 43*, 1420-1434.

Chitra, K., & Subashini, B. (2013). Data Mining Techniques and its applications in banking sector. *International Journal of Emerging Technology and Advanced Engineering*, 3(8), 219-226.

Coe, J. (2004a). Segmentation for communications. *In The fundamentals of business to business sales and marketing* (pp. 71-94). New York: McGraw-Hill.

Coe, J. (2004b). The integration of direct marketing and field sales to form a new B2B sales coverage model. *Journal of Interactive Marketing, 18*, 62-77.

Compton, J. (2012). *So many channels, so few quality leads.* Retrieved from DMNews Data. Strategy. Technology.: www.dmnews.com

Coppock, D. (2002). Why Lift? - Data Modeling and Mining. *Information Management Online*.

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning - ICML (pp. 233-240). Pittsburgh.

Dean, J. (2014). *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners.* New Jersey: Wiley.

D'Haen, J., & Poel, D. V. (2013). Model-supported business-to-business prospect prediction based on an iterative customer acquisition. *Industrial Marketing Management 42*, 544-551.

Du, W., & Zhan, Z. (2002). Building Decision Tree Classifier on Private Data. Proceedings of the International Conference on Privacy, Security and Data Mining, Volume 14 (pp. 1-8). IEEE.

Eggert, A., & Serdaroglu, M. (2011). Exploring the impact of sales technology on sales-person performance: A task-based approach. *Journal of Marketing Theory & Practice, 19*, 169-186.

Gillin, P., & Schwartzman, E. (2011). Lead Generation. In *Social Marketing to the Business Customer: Listen to your B2B Market, generate major account leads, and build client relationships* (pp. 156-175). Hoboken, NJ: Wiley.

Goldie, L. (2007). *Digital marketers see online lead generation as a major growth area.* United Kingdom: New Media Age.

Gordon, S. (2018). How video can enhance prospecting and increase lead converisons - tips to expedite the speed of your sales funnel. *HCM Sales, Marketing & Alliance Excellence*, 13-15.

Greenyer, A. (2000). The use of learning classifier system in the direct marketing industry. C. Brebbia, & N. Ebecken, Eds., *Data Mining II*.

Guifang, G., & Youshi, H. (2010). Research on the Application of Data Mining to Customer Relationship Management in the Mobile Communication Industry, 597-599.

Guo, F., & Qin, H. (2017). Data Mining Techniques for Customer Relationship Management. *IOP Conf. Series: Journal of Physics, 910*. doi:10.1088/1742-6596/910/1/012021

Gupta, N., Wasid, M., & Ali, R. (2016). Analysis of Complex Data in Telecommunications Industry. Proceedings of the International Conference on Computer and Information Technology (pp. 1-4). IEEE.

Han, J., & Kamber, M. (2001). Data Mining: Concepts and Techniques. *Academic Press*.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques* (3rd ed.). Elsevier Inc.

Hand, D., Mannila, H., & Smyth, P. (2001). Principles of Data Mining. MIT Press.

Hosmer, D. W., & Lemeshow, S. (1989). Applied Logistic Regression. The Multuple Logistic Regression Model, 25-37.

Hu, X. (2001). Using rough sets theory and database operations to construct a good ensemble of classifiers for data mining application. Proceedings of the International Conference on Data Mining (pp. 233-240). IEEE.

Hu, X. (2002). Comparison of Classification Methods for Customer Attrition Analysis. Proceedings of the 3rd International Conference on Rough Sets and Current Trends in Computing (pp. 487-492). Springer.

Hu, X. (2005). A Data Mining Approach for Retailing Bank Customer Attrition Analysis. *Applied Intelligence 22*, 47-60.

Kim, S. (2007). Relational Behaviors in Marketing Channel Relationships: Transaction cost implications . *Journal of Business Research, 60(11)*, 1125 - 1134.

Kohavi, R., & Provost, F. (1998). Glossary of Terms. Machine Learning - Special Issue on Applications of Machine Learning and the Knowledge Discovery Process. *Machine Learning, 30*, 271-274.

Krozel, K. E. (2019). Solving the Client Acquisition Equation: Digital Marketing Lead Generation. *LIMRA's MarketFacts*, 66-71.

Kumar, G., Kongara, V., & Ramachandra, G. (2013). An efficient ensemble based classification techniques for medical diagnosis. *International Journal of Latest Technology in Engineering, Management & Applied Science*, 5-9.

Lian, Y., Wolniewicz, R., & Dodier, R. (2004). Predicting Customer Behaviour in Telecommunications. *Intelligent Systems, 19*(2), 50-58.

Lin, Q., Wan, Y., & Pu, F. (2010). *Study on Customer Loyalty and Influential Factors in Mobile Telecommunications.* Beijing University of Posts and Telecommunications, School of Economics and Management. China: IEEE.

Ling, X., & Li, C. (1998). Data Mining for Direct Marketing: Problems and Solutions. AAAI Press. In Proceedings of the 4th KDD Conference (pp. 73-79).

Linoff, G. S., & Berry, M. (2011). *Data Mining Techniques: For Marketing Sales, and Customer Relationship Management* (3rd ed.). Indianapolis, Indiana: Wiley Publishing Inc.

Lu, N., Lin, H., Lu, J., & Zhang, G. (2014, May). A Customer Churn Prediction Model in Telcom Industry Using Boosting. *Transactions on Industrial Informatics, 10, No 2*, 1659 - 1665.

Mehrotra, A., & Agarwal, R. (2009). Classifying customers on the basis of their attitudes towards telemarketing. *Journal of Targeting, Measurement and Analysis for Marketing, 17*, 171-193.

Metzger, M. (2005). Using water testing to convert prospects into leads and leads into customers. *WC&P International, 47*, 7-8.

Monat, J. (2011). Industrial sales lead conversion modeling. . *Marketing Intelligence & Planning, 29*, 178-194.

Moody, D., & Walsh, P. (1999). Measuring The Value of Information: an asset valuation approach. *Seventh European Conference of Information Systems.* Copenhagen.

Moro, S., Laureano, R. M., & Cortez, P. (2011). *Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology.*

Murthy, S. K. (1998, December). Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. *Data Mining and Knowledge Discovery, 2*, 345-389.

Neville, P. (1999). Decision Trees for Predictive Modeling.

Olson, D. L., & Delen, D. (2008). *Advanced Data Mining Techniques. Data Mining Processes.* Springer.

Olson, J. E. (2003). The Data Quality Problem. In Elsevier (Ed.), *Data Quality: The Accuracy Dimension* (pp. 1-15). Morgan Kaufmann Publishers.

O'Reilly, C. (1980). Individuals and Information Overload in Organizations: Is More Necessarily Better? *Academy of Management Journal, 23,(4).*

Page, C., & Luding, Y. (2003). Bank Manager's Direct Marketing Dilemmas - Customer's Attitudes and Purchase Intention. *International Journal of Bank Marketing*, 21(3), 147 - 163.

Pallegedara, A., Amaratunga, V., Gopura, R., & Jayathileka, P. (2006). AI Based Approach of Predicting the Credit Limits of Users to Middle Customer Based Mobile Communication Services. Proceedings of the 1st International Conference on Industrial and Information Systems (pp. 588-592). Sri Lanka: IEEE.

Patterson, L. (2007). Marketing and Sales alignment for improved effectivess. *Journal of Digital Asset Management, 3*, 2592-2602.

Rahm, E., & Hai Do, H. (2000). Data Cleaning: Problems and Current Approaches. *Bulletin of the Technical Committee on Data Engineering* (pp. 1-13). IEEE Computer Society.

Richards, K., & Jones, E. (2008). Customer Relationship Management - Finding value drivers. *Industrial Marketing Management, 37*, 120-130.

Samuels, C. (2013). Generating an unlimited amount of leads for your business. In *10 Lead Generation and Marketing Strategies* (pp. 49 - 62). Inspired Business Services .

Shaoling, D., & Yan, L. (2008). A CRM Model Based on Mining Unstructured Customers' Data. IEEE.

Silverstein, B. (2012). *Sales Leads 123: Generating, Qualifying and Converting Sales Leads in 3 proven steps.* Smashwords Edition.

Velocify. (2012). *The Ultimate Contact Strategy: How to use phone and email for contact and conversion success.* Velocify.

Wang, R., & Strong, D. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems, 12*.

Wang, Y., Sanguansintukul, S., & Lursinsap, C. (2008). The Customer Lifetime Value Prediction in Mobile Telecommunications. Proceedings of the 4th IEEE International Conference on Management of Innovation and Technology (pp. 565 - 569). Bangkok, Thailand: IEEE.

Willis, C. J., & Flo, T. R. (2016). Cleaning up the Lead Generation Highway - the newly-regulated world of online lead generation. *Journal of Internet Law*, 3-7.

Witten, I., & Frank, E. (2005). *Data Mining - Pratical Machine Learning Tools and Techniques* (2nd ed.). USA: Elsevier.

Yang, X., Zhang, X., & Zuo, F. (2008). Who Talk More? The Impact of Relational Attributes on Mobile Phone Users' Communication Behavior. Proceedings of the International Seminar on Business and Information Management (pp. 67 - 70). China: IEEE.

Yu, Y., & Cai, S. (2007). A new approach to customer targeting under conditions of information shortage. *Marketing Intelligence & Planning, 25*, 343-359.

Zhang, X.-h., Yang, X.-c., Shi, W.-h., & Lu, T.-j. (2008). Data Mining Based Marketing Support System for Telcom Operators. IEEE.

Zhao, C., Wu, Y., & Gao, H. (2008). Study on Knowledge Acquisition of the Telcom Customers' Consuming Behaviour Based on Data Mining. IEEE.

Zhao, H., Zhang, X.-h., Wang, Q., Zhang, Z.-c., & Wang, C.-y. (2014). Customer Segmentation on Mobile Online Behavior. Proceedings of the 21th International Conference on Management Science & Engineering (pp. 103-109). Helsinki, Finland: IEEE.

ANEXOS

Anexo A – Comparação gráfica da Curva ROC dos modelos segundo os vários métodos de seleção de variáveis.

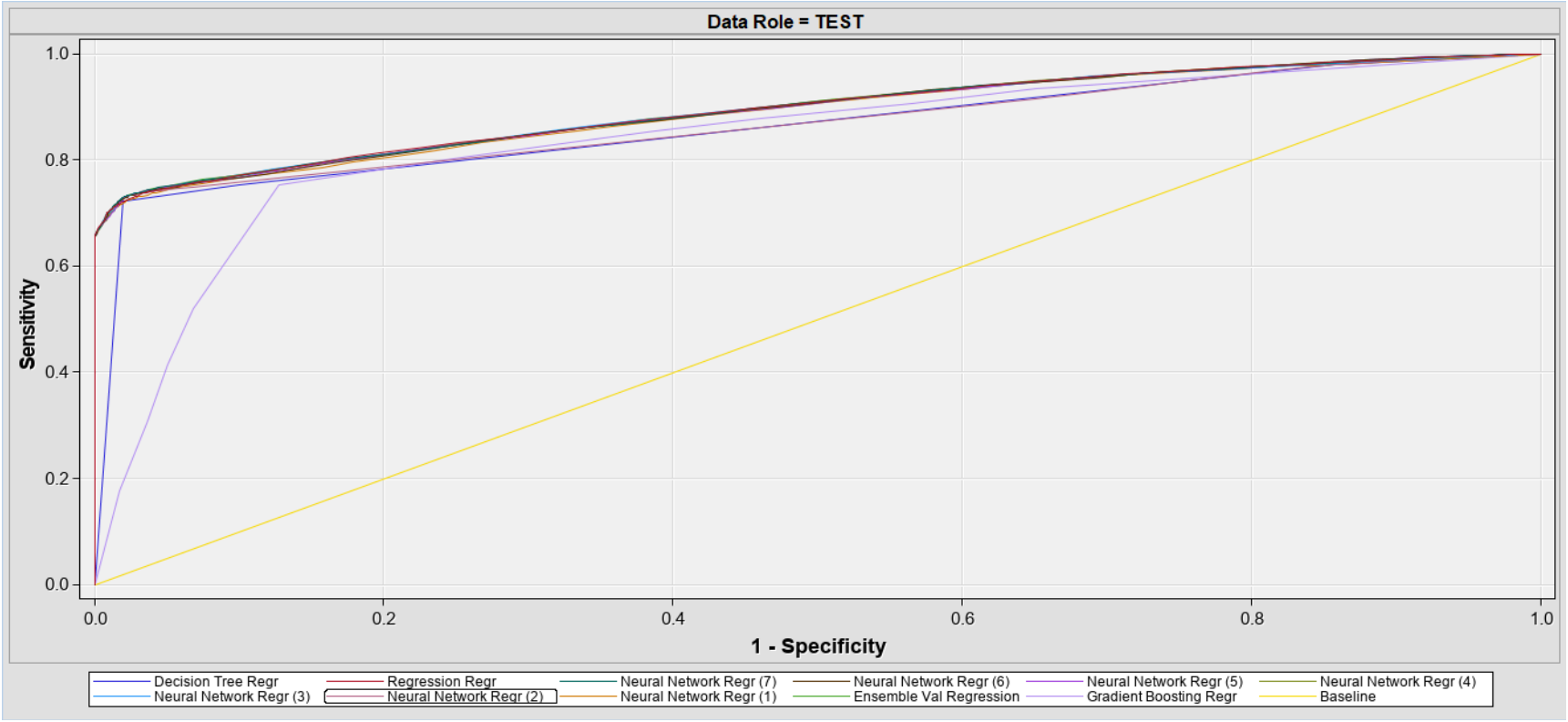

Figura A.1 - Comparação da Curva ROC dos modelos segundo o método de seleção Regressão (Stepwise), para o conjunto de dados de teste.

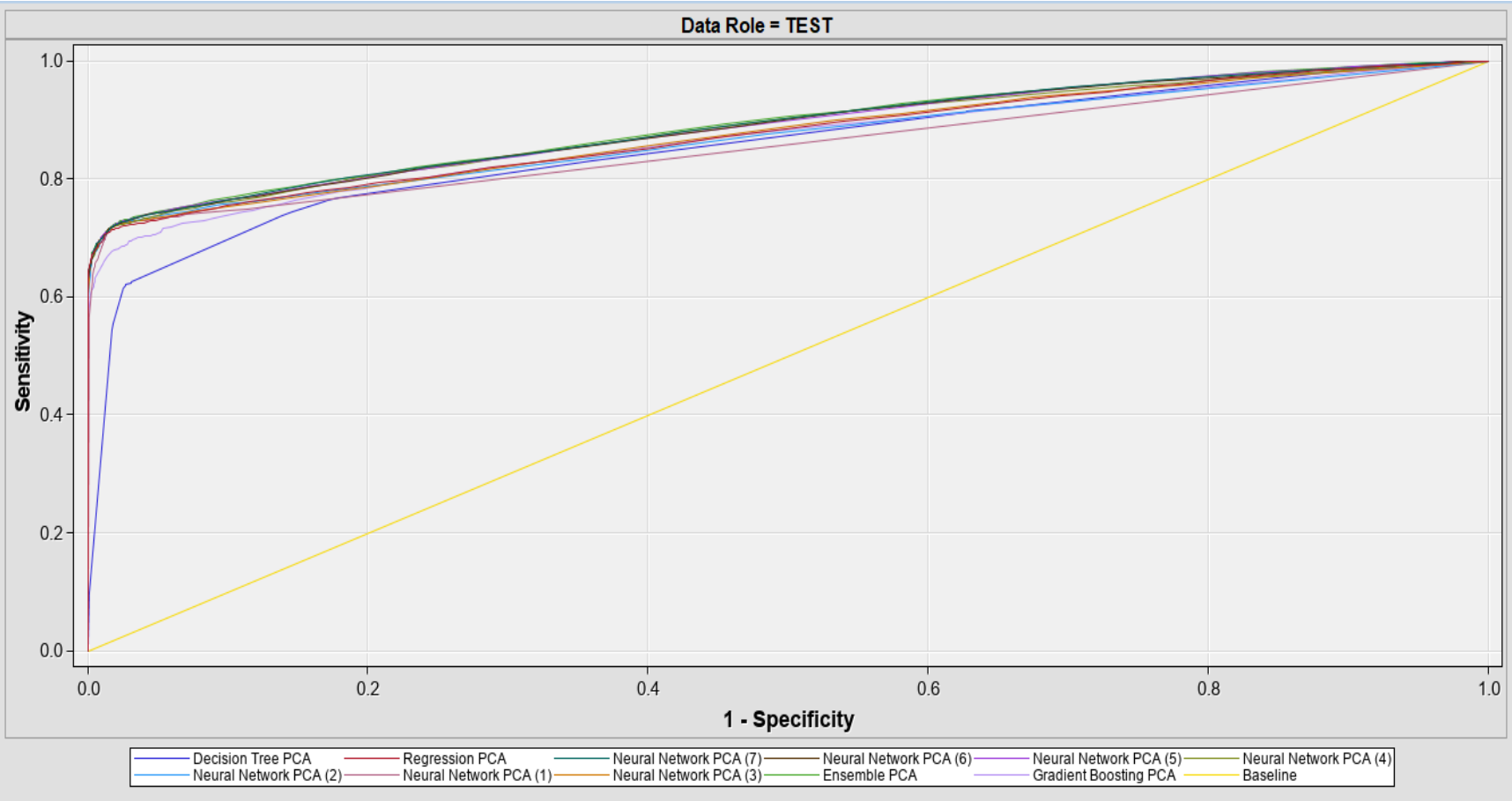

Figura A.2 - Comparação da Curva ROC dos modelos segundo o método de seleção PCA, para o conjunto de dados de teste.

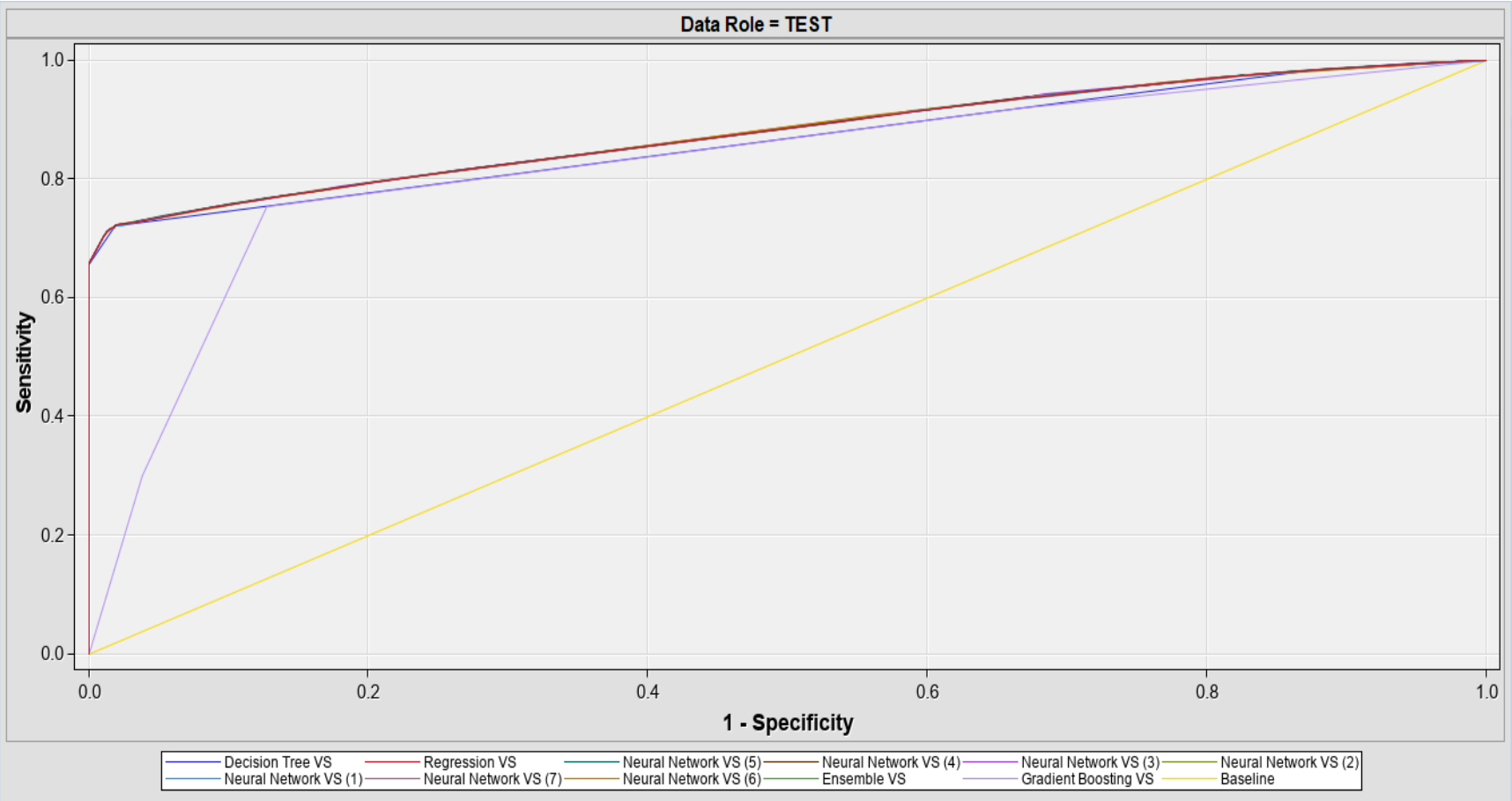

Figura A.3 - Comparação da Curva ROC dos modelos segundo o método de seleção *Variable Selection*, para o conjunto de dados de teste.

Anexo B – Comparação estatística de modelos segundo os vários métodos de seleção de variáveis.

Tabela B.10– Comparação estatística de modelos segundo o método de seleção Regressão (*Stepwise)*, para o conjunto de dados de teste.

Regressão (*Stepwise*)

Modelo	Taxa de Erro	Modelo	Índice ROC	Modelo	Coeficiente de Gini	Modelo	Estatística de Kolmogorov-Smirnov
Neural Network 7	0,1615	Ensemble	0,894	Ensemble	0,788	Ensemble	0,710
Ensemble	0,1618	Neural Network 7	0,894	Neural Network 7	0,787	Neural Network 5	0,710
Neural Network 5	0,1619	Regression	0,894	Neural Network 3	0,787	Neural Network 7	0,710
Neural Network 3	0,1620	Neural Network 5	0,893	Neural Network 4	0,787	Neural Network 2	0,709
Neural Network 2	0,1625	Neural Network 3	0,893	Regression	0,787	Neural Network 3	0,709
Neural Network 4	0,1625	Neural Network 4	0,893	Neural Network 5	0,785	Neural Network 4	0,709
Neural Network 6	0,1631	Neural Network 6	0,893	Neural Network 6	0,785	Neural Network 6	0,709
Regression	0,1647	Neural Network 1	0,89	Neural Network 1	0,78	Regression	0,703
Decision Tree	0,1664	Neural Network 2	0,875	Neural Network 2	0,751	Decision Tree	0,702
Neural Network 1	0,1667	Decision Tree	0,863	Decision Tree	0,727	Neural Network 1	0,701
Gradient Boosting	0,1953	Gradient Boosting	0,839	Gradient Boosting	0,679	Gradient Boosting	0,625

Tabela B.2 – Comparação estatística de modelos segundo o método de seleção Análise de Componentes Principais, para o conjunto de dados de teste.

Análise de Componentes Principais							
Modelo	**Taxa de Erro**	**Modelo**	**Índice ROC**	**Modelo**	**Coeficiente de Gini**	**Modelo**	**Estatística de Kolmogorov-Smirnov**
Ensemble	0,1644	Ensemble	0,890	Ensemble	0,780	Ensemble	0,705
Neural Network 5	0,1649	Neural Network 7	0,888	Neural Network 7	0,777	Neural Network 5	0,704
Neural Network 7	0,1653	Neural Network 5	0,887	Neural Network 5	0,774	Neural Network 7	0,702
Neural Network 6	0,1663	Neural Network 6	0,887	Neural Network 6	0,774	Neural Network 6	0,701
Neural Network 2	0,1674	Neural Network 4	0,886	Neural Network 4	0,771	Neural Network 2	0,701
Neural Network 3	0,1675	Neural Network 3	0,877	Neural Network 3	0,753	Neural Network 4	0,701
Neural Network 1	0,1679	Regression	0,876	Regression	0,753	Neural Network 3	0,700
Neural Network 4	0,1691	Neural Network 2	0,873	Neural Network 2	0,746	Neural Network 1	0,700
Regression	0,1767	Gradient Boosting	0,868	Gradient Boosting	0,737	Regression	0,695
Gradient Boosting	0,1854	Neural Network 1	0,866	Neural Network 1	0,732	Gradient Boosting	0,665
Decision Tree	0,2079	Decision Tree	0,853	Decision Tree	0,705	Decision Tree	0,599

Tabela B.3 – Comparação estatística de modelos segundo o método de *seleção Variable Selection, para* o conjunto de dados de teste.

Variable Selection

Modelo	Taxa de Erro	Modelo	Índice ROC	Modelo	Coeficiente de Gini	Modelo	Estatística de Kolmogorov-Smirnov
Regression	0,1664	Ensemble	0,878	Neural Network 3	0,756	Ensemble	0,702
Neural Network 5	0,1664	Neural Network 6	0,878	Neural Network 4	0,756	Decision Tree	0,702
Decision Tree	0,1664	Neural Network 7	0,878	Neural Network 5	0,756	Neural Network 6	0,702
Neural Network 4	0,1664	Neural Network 3	0,878	Ensemble	0,755	Neural Network 7	0,702
Ensemble	0,1665	Neural Network 4	0,878	Neural Network 6	0,755	Neural Network 3	0,702
Neural Network 6	0,1665	Neural Network 5	0,878	Neural Network 7	0,755	Neural Network 4	0,702
Neural Network 7	0,1665	Neural Network 2	0,878	Regression	0,755	Regression	0,702
Neural Network 3	0,1665	Regression	0,877	Neural Network 2	0,755	Neural Network 5	0,702
Neural Network 2	0,1665	Neural Network 1	0,877	Neural Network 1	0,755	Neural Network 2	0,702
Neural Network 1	0,1665	Decision Tree	0,872	Decision Tree	0,744	Neural Network 1	0,702
Gradient Boosting	0,1955	Gradient Boosting	0,823	Gradient Boosting	0,646	Gradient Boosting	0,625

Anexo C – Modelo Preditivo para classificação de novas *Leads*, construído através do *SAS Enterprise Miner*.